

**Summary Report (Draft)
External Peer Review of FDA's**

***Draft Risk Ranking Model for Product Tracing as Required by Section
204 of FSMA – Model Review***

Contract No. HHSF223201210011B
BPA No. 05

March 8, 2016

Prepared for:

Lori Papadakis & Yuhuan Chen, Ph.D.
U.S. Food and Drug Administration
Center for Food Safety and Applied Nutrition
5100 Paint Branch Parkway
College Park, MD 20740

Prepared by:

Versar, Inc.
6850 Versar Center
Springfield, VA 22151

Peer Reviewers:

Panos G. Georgopoulos, Ph.D.
Igor Linkov, Ph.D.
Thomas Ross, Ph.D.
Moez Sanaa, DVM, MSc, Ph.D.
Nga A. Tran, Dr.P.H., M.P.H., CIH

TABLE OF CONTENTS

I.	INTRODUCTION	1
II.	CHARGE TO REVIEWERS	2
III.	INDIVIDUAL REVIEWER COMMENTS	4
	Reviewer #1	5
	Reviewer #2	16
	Reviewer #3	29
	Reviewer #4	36
	Reviewer #5	57
IV.	PEER REVIEWER COMMENT TABLE.....	62
	I. General Impressions	63
	II. Response to Charge Questions.....	65
	III. Specific Observations on <i>Draft Report for Peer Review: Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA (RRM-PT Draft Report)</i> within the context of the model itself.....	113
	IV. Specific Observations on Appendices to the Draft Report for Peer Review: Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA (RRM-PT Draft Report) within the context of the model itself.....	116
	V. Specific Observations on Risk Ranking Model for Product Tracing: User’s Guide Short Version	116
	VI. Specific Observations - Provide specific observations, corrections, or comments on the Access Database FDA’s High Risk Foods (HRF) Model.....	117

I. INTRODUCTION

Section 204(d) (2) of the FDA Food Safety Modernization Act (FSMA) requires FDA to designate high risk foods (HRFs) for which additional recordkeeping requirements are necessary to protect public health. Under FSMA, FDA's designation of HRFs is based on six factors. Although FSMA Section 204 requires FDA to designate "high risk foods," in order to apply the FSMA factors, it is necessary to first take into account characteristics of both foods and hazards, i.e., food-hazard pairs. To address the requirements of FSMA Section 204, CFSAN has developed a data-driven model, the *Draft Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA*, that uses seven explicit criteria related to public health risk. Both microbial and chemical hazards are considered in the model as required by FSMA in the HRF designation.

For this peer review, five experts were selected to answer 12 charge questions and to evaluate and provide written comments on the HRF Model, the HRF Model Report, the HRF Model User's Guide, the HRF Model Design Report, and select data and food-hazard pairs. The peer review focused on the conceptual framework of the HRF model and associated equations and the underlying relational database, as well as selected food-hazard combinations to evaluate the accuracy of scoring and assess the usability of its interactive interface.

Peer Reviewers:

Panos G. Georgopoulos, Ph.D.

Environmental and Occupational Health Sciences Institute (EOHSI)
Piscataway, NJ 08854

Igor Linkov, Ph.D.

Environmental Laboratory
U.S. Army Corps of Engineers
Boston, MA 02130

Thomas Ross, Ph.D.

School of Land and Food – Tasmanian Institute of Agriculture
University of Tasmania
Private Bag 54, Hobart, TASMANIA 7001 AUSTRALIA

Moez Sanaa, DVM, MSc., PhD

French Agency for Food, Environmental and Occupational Health & Safety
Risk Assessment Department
94701 MAISONS-ALFORT Cedex
France

Nga A. Tran, Dr.P.H., M.P.H., CIH

Exponent, Inc.
Washington, DC 20036

II. CHARGE TO REVIEWERS

FDA has developed a draft risk ranking model to inform the designation of high-risk foods for which additional recordkeeping requirements are appropriate and necessary to assist product tracing, as required under Section 204(d)(2) of the FDA Food Safety Modernization Act (FSMA). The draft model was developed through an iterative process that involved, among other things, using the FSMA statutory factors to define criteria and scoring functions of the criteria, and collecting data relevant to the scoring criteria for food-hazard pairs to identify those foods which should be designated as high-risk for future consideration in policy decision.

The focus of this review is on the model, in the context of the overall risk ranking approach, criteria and results. Note: a separate panel is reviewing the underlying data.

Charge Questions:

1. The draft risk ranking model uses the FSMA statutory factors to define seven criteria for scoring food-hazard pairs.
 - a. Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.
 - b. Within the bounds of the FSMA-mandated factors, are there additional criteria beyond the seven criteria that should be considered? If so, please describe these additional criteria that might be considered and why.
2. Are the scoring definitions for all criteria appropriate?
 - a. Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and undeclared allergens)? If not, please describe changes that might be considered and why.
 - b. Is the value function 0-1-3-9 and scoring matrix appropriate for the intended purpose to inform the designation of high-risk foods? If not, please describe changes that might be considered and why.
3. Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.
4. Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.
5. Considering the various scenarios (described in Section 7.3 and Appendix N) to aggregate food-hazard pairs in order to identify the foods which should be identified as high-risk vs. not high-risk, which option(s) are more appropriate to consider and why? Are there additional aggregation method(s) that might be considered? Please explain.
6. Given the underlying data supporting the scoring, what are the considerations to take into account when identifying high risk vs. not high risk food-hazard pairs or foods? What are the

- pros and cons in establishing a threshold considering all three types of hazards in the model, vs. drawing a line separately for microbial, chemical and undeclared allergens in foods?
7. Are the seven criteria and scoring definitions implemented appropriately in the SAS codes (Appendix P) and the Access Model?
 - a. Does the scoring logic described in Section 4 Figures 4-1 to 4-7) appropriately represent the scoring definitions described for each of the criteria in Section 3? If not, please describe what changes need to be made to correct it.
 - b. Are the scoring logic and order of preference accurately implemented in the SAS codes? (Please select 2-3 out of the 7 criteria for this evaluation). If not, please specify what changes need to be made.
 - c. Are equations 1 through 4 and data weighting factors accurately implemented in the model (either the SAS codes or the Access Model)?
 8. In the Access Model, is the underlying relational database including lookup tables and algorithm appropriately designed and implemented? If not, please explain what changes should be considered.
 9. Is the user interface of the Access Model sufficiently described for the user to understand each component of the model, e.g., foods, hazards, ranking criteria, results, and cited references?
 10. How often should the model be updated, considering the data sources and data currently available and types of data that might become available in the future?
 11. Is the draft report clear in its description of the risk ranking approach, criteria and scoring definitions, and model limitations? If not, please identify which aspects are unclear or could be more transparent.
 12. Do you have any additional comments? Please share them in your review.

III. INDIVIDUAL REVIEWER COMMENTS

Reviewer #1

Peer Review Comments on FDA's *Draft Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA:*

HRF Model, the HRF Model Report, the HRF Model User's Guide, and
the HRF Model Design Report

Reviewer #1

I. GENERAL IMPRESSIONS

Despite the complex nature of connecting multiple factors and a myriad of data sources to construct the high risk foods model, the current draft model is straightforward, easily understood and generally logical. However, the simplicity of the model with seven criteria that are proxy for potential exposure and risk of illnesses from “high risk foods”, necessitates rigorous scrutiny of the underlying data and process that were relied upon to inform the scoring of each of the seven criteria for each food-hazard pair, as well as how the seven criteria are combined and how food-hazard pairs are integrated to index “high risk foods”. It is this reviewer's understanding that examination of the underlying data is being undertaken by another peer review panel, hence the comments provided herein focus on the seven criteria scoring method and integration of the seven criteria scores within a food-hazard pair as well as integration across multiple hazards for a given food. Overall, it is this reviewer's impression that scoring and underlying data supporting scoring for each of the seven criteria have been thoroughly examined by the Agency in the development of the current model. However, the process of integrating scores for each food-hazard pair and integrating across hazards for each food is still work in progress, and further sensitive assessment is needed to assure that the method selected for the final model is sound and supported by the most robust science/policy rationale. It is also noted that significant progress has been made to adequately capture chemical risks in this current draft model. Some degree of unbalanced emphasis on microbial risks remain and are noted below in response to the various charge questions of this peer review.

II. RESPONSE TO CHARGE QUESTIONS

1. *The draft risk ranking model uses the FSMA statutory factors to define seven criteria for scoring food-hazard pairs.*

a. *Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.*

The seven criteria in the draft model are appropriate for a multi-criteria decision approach. These criteria capture the essence of the FSMA factors; each FSMA factor was represented by two criteria, except for FSMA (iii) and (iv) which were represented by a single criterion 5. The seven criteria are proxy for exposure and risks, hence, appropriate for the purpose of a risk-based ranking model tool.

The emphasis of FSMA on manufacturing aspects (two FSMA factors), but represented by just one criterion 5 should be noted and considered in the weighting of the seven criteria in the aggregation across food-hazard pairs to derive a composite score of a single food (more later).

- (i) the known safety risks of a particular food, including the history and severity of foodborne illness outbreaks attributed to such food, taking into consideration

foodborne illness data collected by the Centers for Disease Control and Prevention;

- criteria 1 (frequency and occurrence of outbreaks)
 - criteria 2 (severity of illness)
- (ii) the likelihood that a particular food has a high potential risk for microbiological or chemical contamination or would support the growth of pathogenic microorganisms due to the nature of the food or the processes used to produce such food;
- criteria 3 (likelihood of contamination)
 - criteria 4 (growth potential)
- (iii) the point in the manufacturing process of the food where contamination is most likely to occur;
- criteria 5 (manufacturing process/contamination intervention)
- (iv) the likelihood of contamination and steps taken during the manufacturing process to reduce the possibility of contamination
- criteria 5 (manufacturing process/contamination intervention)
- (v) the likelihood that consuming a particular food will result in a foodborne illness due to contamination of the food;
- criteria 6 (consumption)
 - criteria 3 (likelihood of contamination)
- (vi) the likely or known severity, including health and economic impacts, of a foodborne illness attributed to a particular food.
- criteria 7 (economic impact)
 - criteria 2 (severity of illness)
 -
- b. ***Within the bounds of the FSMA-mandated factors, are there additional criteria beyond the seven criteria that should be considered? If so, please describe these additional criteria that might be considered and why.***

The seven criteria specified in the draft model appear to cover the FSMA factors – see above.

2. Are the scoring definitions for all criteria appropriate?

The scoring definitions for the seven criteria are generally appropriate for microbial risks and undeclared allergens. Some scoring definitions for chemicals are difficult to follow and commented further below.

- a. ***Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and undeclared allergens)? If not, please describe changes that might be considered and why.***

Criterion 1 (frequency of outbreaks and occurrence of illness)

For chemicals, it is said that scoring is based on expert elicitations and the definition for the scores (1, 3, and 9) uses terms “little, some, and compelling” evidence. What constitute little, some and compelling? Was there a rigorous evidence-based approach with guiding principal that was applied for consistency in reaching these ratings? Further by examining the list of experts from which the scores were elicited, there are 3 experts labelled as “toxicologists”. Given the range of chemicals potentially involved, such elicitation process would benefit from a larger pool of toxicologists.

Per appendix K, Table 5, there are 160 cases where expert elicitation was used to assign scores to criterion 1 for chemicals, 33 for undeclared allergens, and none for microbes. Given the emphasis of the expert opinion on the value/score of criterion 1 for chemicals, a more robust elicitation (i.e., larger pool of experts) to capture range of expert opinions would be warranted.

Also for undeclared allergens, appendix K, page 7, noted that scores were elicited from an allergens expert separately from the microbial and chemical hazard groups. However, from the list of experts (Table 4-3 in draft report, or Table 2 of appendix K), there are no allergen experts. For transparency, the allergen experts should be provided in the draft report.

Criterion 2 (severity of illness)

For chemicals it is said that the definition for scoring severity with acute exposure is based out of ICMSF (2001) and used for scoring in the draft risk model, according to the definitions in Table 2.1. Similarly, for chronic the definitions in Table 2.2 are applied for scoring. Further into Section 4.2.2 of the draft report, it appears that the scoring was done by subject matter experts using the definitions in 2.1 and 2.2. Who are these subject matter experts? Was this done through the same expert elicitation process as described in Appendix K? Per appendix K, there are 25 cases where expert elicitation was used but there are more than 25 cases of food-chemical pairs in the draft model. More transparency is needed in derivation of the scores for criterion 2 for chemicals.

Criterion 3 (likelihood of contamination)

For chemical hazards, this likelihood of contamination is said to be determined based on percent positive above action levels or allowable levels. Are the weights ($n*gw*dw$) that is applicable in the case of microbes also applicable to chemicals (and allergens)? They should be if they are not already.

The USDA Agricultural Marketing Service (AMS) Pesticide Data Program (PDP) is a potential data source to rely upon for scoring this criterion. Is there a reason why it was not included as a reference source?

Criterion 4 (growth potential /shelf life)

This criterion is strictly for microbes. The current method of aggregation of scores across seven criteria to derive the composite score for each food-hazard pair by summing (equal weights to all seven criteria) is preferentially selecting food-microbe pairs (i.e., forcing higher ranks on these pairs over chemical and undeclared allergens where there is usually no growth).

To avoid the current imbalance, both criteria 3 and 4 are relevant to address FSMA factor ii (i.e., “the likelihood that a particular food has a high potential risk for microbiological or chemical contamination *or* would support the growth of pathogenic microorganisms due to the nature of the food or the processes used to produce such food”). For chemical, criterion 3 is relevant to address FSMA ii. For microbial hazard, conceptually the score for FSMA ii can be derived based on the composite of criteria 3 and 4 as followed:

C3 Likelihood of contamination	High (9)	3	9	9
	Medium (3)	1	3	9
	Low (1)	1	1	3
		Low (1)	Medium (3)	High (9)
		C4 Likelihood of growth		

Criterion 6 (consumption)

In the current draft model/report, it is stated that the consideration of both criteria 3 (likelihood of contamination) and C6 (percent consumer) defines the likelihood of consuming a particular contaminated food result in illness. While the percent consumer may be a reasonable proxy for microbial and undeclared allergen risk, for chemical risk, the dose make the poison. Thus, there is a need to know how much is consumed, i.e. the likelihood of consuming a particular contaminated food resulting in illness is a consideration of both criterion 3 (likelihood of contamination) and criterion 6 (percent consumer*amount consumed). Conceptually, scoring for criterion 6 for chemicals may be as followed:

Percent consumers	High	3	9	9
	Medium	1	3	9
	Low	1	1	3
		Low	Medium	Day
		Amount consumed per day (g/day)		

Criteria 7 – economic impact

Scallan et al. (2011) and Palmer et al. (2013) are noted as sources for this information for scoring outbreaks and sporadic cases (page 24 of draft report). What are the reference sources for chemical related endpoints? What are the endpoints that were captured in this metric? Was expert elicitation utilized for missing data for chemical endpoints? If so, were health economists among the experts from which the information was elicited? On page 66 of the draft report, Figure 4-7 indicates that expert judgment is used when no quantitative data are available. Who are these experts? Transparency is needed here. In the example on apple juice-arsenic, page 68, how are the dollar amount assigned to the 51.4 cancer cases/year?

- b. Is the value function 0-1-3-9 and scoring matrix appropriate for the intended purpose to inform the designation of high-risk foods? If not, please describe changes that might be considered and why.***

The sensitive analysis in Appendix N (Section 6. Impact of scoring scale) explored the use of an alternative ordinal scale of 1, 2, 3 and 4 (i.e., Anderson model) and impact on model results. The

output in Table 9 of the Appendix N showed somewhat comparable results for the top 20 commodities when using the 0, 1, 3, and 9 (the current model) and when using the Anderson models, and it was concluded that the scoring scale in the HRF model did not dramatically affect the ranking results. The following suggestions are made to expand the sensitivity analysis to strengthen the current conclusion that ranking of foods is insensitive to the types of value function being used:

- Since there are 335 specific foods in the current model, the table showing the comparability of ranks should be expanded to show comparability of the top 10-20% ranks from each type of ranking.

The comparability of the value function being used to score the seven criteria should also be examined at the food-hazard pair level. Insight on whether a value function has an influence on the rank order of the food-hazard pairs, can help inform their potential impact on the aggregated scores for a food, depending on what method of aggregation across multiple hazards.

3. *Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.*

In the current draft model, the algorithm to generate a risk score for a particular food-hazard pair (FRS) is the sum of the weighted scores for the seven criteria. What was the rationale for addition as the mathematical operation to combine the seven criteria to derive a composite score of each food-hazard pair? Were other options to combine considered and sensitive analyses done? Other options may include weighted average of the seven criteria, weighted product of the seven criteria, or a combination of weighted addition of likelihood for exposure indicators and likelihood of illness indicators and multiplication of these two composite likelihoods (i.e., likelihood of risk = exposure dose x dose-response). At the very least, this should be discussed and rationale provided as to why they were not pursued.

4. *Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.*

In the current draft model, the default weight of 10 is applied to all seven criteria, i.e., no differential weighting, and the scores for the seven criteria are summed to derive FRS for each food-hazard pair.

As noted earlier, the current summation with criterion 4 given full weight, would preferentially rank food-micro hazard pairs higher than chemical and undeclared allergen pairs. Suggestion is provided above to avoid the imbalance between the hazard types.

Since the intent of the model is to address FSMA factors/requirements and the seven criteria that were derived intentionally to capture the FSMA factors, it may be best to focus the weighting of

these seven criteria based on FSMA emphasis. The table below assumed a default equal weight of 10 for each of the six FSMA factors (i.e., FSMA weight), hence a total weight of 60. The table also summarized the seven criteria in the draft model per each FSMA factor and the FSMA weight for each of the criterion is derived by dividing the FSMA default weight by the number of the model criteria assigned to the FSMA factor. The final weight for each criterion is derived by summing the FSMA weight. Criterion 4 is only applicable to micro hazards so noted with C3 (see above for suggestion to composite with C3 for micro hazard).

FSMA Factor	FSMA Weight	Criteria	FSMA weight	Final criterion weight	
FSMA i	10	C1	5	C1	5
		C2	5	C2	5
FSMA ii	10	C3	5	C3 (&C4)	15
		C4	5	C5	20
FSMA v	10	C3	5	C6	5
		C6	5	C7	10
FSMA iii	10	C5	10		
FSMA iv	10	C5	10		
FSMA vi	10	C7	10		
Total	60		60		

5. Considering the various scenarios (described in Section 7.3 and Appendix N) to aggregate food-hazard pairs in order to identify the foods which should be identified as high-risk vs. not high-risk, which option(s) are more appropriate to consider and why?

What constitute/define “high-risk food” would ultimately have an impact on what aggregation approach is used to derive composite scores for foods for rank ordering. High risk could be based on: 1) frequency of hazards in a single food (i.e., no. of hazards in a given food), or 2) when hazard occurs in a food it is a very high risk (i.e., high FRS score), or 3) a combination of both (1) and (2).

In the various analyses provided in the draft report Table 7-4 as well as in appendix N, FDA attempted to tackle these definitions by various aggregation schemes, i.e., summation, average, maximum FRS, or cutoff based on FRS score and sum, etc. Overall, the current approach as described in Section 7, which is a cutoff for FRS at 270 prior to summing to derive composite score for foods and then rank, is defining “high-risk” food based on (2), i.e., when hazard occurs in a food it is a very high risk (high FRS score above a set cutoff). This approach deliberately dismisses foods with multiple hazards but not of high FRS. While it is not unreasonable to not allow lower risk hazard to have high influence, it is possible that foods with multiple small problems could be considered “high risk” food and this approach could be viewed negatively by some.

a. Are there additional aggregation method(s) that might be considered? Please explain.

Some thoughts should be given further into the cutoff point for the FRS. Various cutoff point should be considered in a sensitivity analysis. If microbial, chemical, and allergens are treated

equally in the raw scores for each criterion, then there is no need for a different cutoff point. However, if the current unbalance score is kept (i.e., microbes have an additional C4 score of up to 9), then a different cutoff (lower) would have to be considered for chemical and allergens.

6. Given the underlying data supporting the scoring, what are the considerations to take into account when identifying high risk vs. not high risk food-hazard pairs or foods?

Based on the appendix J, the percent consumer is based on the entire population. Considerations should be made to incorporate this into criterion 6 to consider sensitive subpopulations such as children, pregnant women, and the elderly.

Also for the economic criterion 7, some discussions, considerations with respects to economic burden to society when the effect is on child/fetus should be provided (it is unclear if these considerations were accounted for in the scoring of this metric in the current model).

The uncertainty/confidence in the scores are tracked. However, it is unclear as to how this information will be used in ranking of high risk foods. There need to be some discussions of what the Agency intends to do with areas where there is lack of confidence of the data and results are highly uncertain.

- a. What are the pros and cons in establishing a threshold considering all three types of hazards in the model, vs. drawing a line separately for microbial, chemical and undeclared allergens in foods?**

See comment above in question 5.

7. Are the seven criteria and scoring definitions implemented appropriately in the SAS codes (Appendix P) and the Access Model?

It should be noted that I am not a SAS programmer and my SAS experience is limited and dated. With this in mind, I reviewed the appendix P as requested and provided some comments below.

- a. Does the scoring logic described in Section 4 Figures 4-1 to 4-7) appropriately represent the scoring definitions described for each of the criteria in Section 3? If not, please describe what changes need to be made to correct it.**

Yes, the scoring logic in Section 4 (Figures 4-1 to 4.7) appropriately represents the scoring definitions. Some questions/issues are provided above under question 2.

- b. Are the scoring logic and order of preference accurately implemented in the SAS codes? (Please select 2-3 out of the 7 criteria for this evaluation). If not, please specify what changes need to be made.**

Criterion 1: appears to correctly implement the scoring of this criterion as described in Section 4, Figure 4-1

Criterion 2: appears to correctly implement the upper portion of Figure 4-2 and Table 2-1. This reviewer cannot see where Table 2-2 and lower portion of Figure 4-2 (where expert judgment

comes in for chemical and allergens). There are codes written for hazard id 73 (methanol), 2, and 3, but it is unclear why.

Criterion 6: appears to be correctly implemented to tabulate percent consumers among U.S. population.

c. Are equations 1 through 4 and data weighting factors accurately implemented in the model (either the SAS codes or the Access Model)?

Equations 1 and 2 (for criterion 3) appear to correctly reflect the data weighting description in the draft report. No SAS codes can be found in appendix P for equations 3 and 4.

8. In the Access Model, is the underlying relational database including lookup tables and algorithm appropriately designed and implemented? If not, please explain what changes should be considered.

The underlying relational database, look-up tables, and algorithm appear to be appropriately designed and implemented.

9. Is the user interface of the Access Model sufficiently described for the user to understand each component of the model, e.g., foods, hazards, ranking criteria, results, and cited references?

The user interface of the model is sufficiently described for this reviewer to understand each component of the model. However, it should be noted that this reviewer is familiar with the model approach so a newcomer may have a different opinion.

10. How often should the model be updated, considering the data sources and data currently available and types of data that might become available in the future?

The frequency of updates of the model should be determined by the underlying data upon which the scores for the seven criteria are developed (e.g., CDC surveillance data release, TDS monitoring data release, NHANES data release, etc.). A 2- to 4-year period for update may be reasonable.

11. Is the draft report clear in its description of the risk ranking approach, criteria and scoring definitions, and model limitations? If not, please identify which aspects are unclear or could be more transparent.

This is a complex process with many layers of data aggregation. The information provided in the draft report and associated appendices are well laid out and easy to followed with adequate details in most areas. The Figure 2-1 demonstrating the relationship between the FSMA factors and the criteria in the high risk food model (HRFM) is helpful to orient the model in context of FSMA. The descriptions and scoring of the seven criteria as described in Section 2 of the current draft is mostly clear and reflective of what the current model intends to do. Section 4.1 and data indicator in Table 4-2 are useful, allowing for a quick understanding of underlying data/metric that are relied upon as proxy for exposure and risk for the food-hazard pairs. The scoring process flow charts for each criterion in the HRFM supplement the description of the seven criteria in

Section 2 and further clarify the current model. The descriptions of the expert elicitation and information in appendix K are helpful allowing readers a better understanding of the sources and limitations associated with the information obtained from this process to fill in data gaps. The description of risk and uncertainty scores is easily understood. Overall the report did a very good job of explaining what has been done and the elements of the model.

12. Do you have any additional comments? Please share them in your review.

None.

III. SPECIFIC OBSERVATIONS ON DRAFT REPORT FOR PEER REVIEW: RISK RANKING MODEL FOR PRODUCT TRACING AS REQUIRED BY SECTION 204 OF FSMA (RRM-PT Draft Report) WITHIN THE CONTEXT OF THE MODEL ITSELF.

Page	Paragraph/Line	Comment
17	Line 19	Table 2-1 title should reflect acute exposure
18	Line 8	Text in [Table 2-2] describing score 9 for chronic health hazards should include examples of endpoints with this type of score
63	Line 1	Examples did not show apple juice and arsenic example
63	Line 36	Salad kit example of no data and expert opinion was used to assign score, who are the experts? Was this from the elicitation process outline in appendix N?
71	Lines 12-13	Who are the allergen experts?
72	Lines 15-17	Who are the FDA subject matter experts? How is this done, is there a report detailing his process? An appendix documenting this process (similar to appendix K for the expert panel elicitation process) would be helpful.
73	Line 10	Why sum? Need to provide rationale
102	Line 27	Appendix L was not provided for peer review

IV. SPECIFIC OBSERVATIONS ON APPENDICES TO THE DRAFT REPORT FOR PEER REVIEW: RISK RANKING MODEL FOR PRODUCT TRACING AS REQUIRED BY SECTION 204 OF FSMA (RRM-PT Draft Report) WITHIN THE CONTEXT OF THE MODEL ITSELF.

Appendix	Page/Row	Paragraph/Line/Column	Comment
K	5	Second paragraph, lines 1 and 2	It is stated that criterion 7 (economic impact) has relatively few data gaps, yet looking at Table 3 on page 6, we see 340 for chemicals. 340 does not appear to be minor data gap here. If the 340 C7 scores for chemicals were based on the expert elicitation as indicated in Table 3 of Appendix K, then who were the experts from which the economic scores elicited from? On the list of experts in Table 2 of appendix K, there is no “health economics” expertise that would be necessary for such an expert elicitation.
N	1, Table 1	Leafy green has FRS of 249	This 249 was said to be based on sum of individual hazard pairs for leafy green in this appendix. However, in draft report on page 102, lines 28-31, it is said to have 12 pairs and

Appendix	Page/Row	Paragraph/Line/Column	Comment
			the risk scores for the pairs range from 130-450. How is it possible that the sum for 12 pairs be less than the score for one pair of 450?
O			Why are these data not included? Is there plan to include?

V. SPECIFIC OBSERVATIONS ON RISK RANKING MODEL FOR PRODUCT TRACING: USER’S GUIDE SHORT VERSION.

Page	Paragraph/Line	Comment
		No comments

VI. SPECIFIC OBSERVATIONS – PROVIDE SPECIFIC OBSERVATIONS, CORRECTIONS, OR COMMENTS ON THE ACCESS DATABASE *FDA’S HIGH RISK FOODS (HRF) MODEL*.

Menu Choice	Tab	Steps taken within the tab	Comment
			No comments

Reviewer #2

Peer Review Comments on FDA's *Draft Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA:*

HRF Model, the HRF Model Report, the HRF Model User's Guide, and
the HRF Model Design Report

Reviewer #2

I. GENERAL IMPRESSIONS

I highly welcome the initiative of FDA to develop the High Risk Foods Model and Associated Food-Hazard Combinations-Model risk ranking model and support the efforts towards a more systematic approach for risk ranking. I wish to congratulate the project team for the huge work done to collect the needed data for 1286 food-hazard combinations. I believe that the initiative that includes both chemical and microbial hazards will be successful and recognized by the other food safety agencies that need a coherent and pragmatic way to rank risk associated with food.

However, as a general remark, I would like to stress the need of better inclusion of uncertainty in the risk ranking process. Because of the high numbers of ranked objects, it is illusory to expect that all the criteria used will be accurately and precisely evaluated for all the food-hazard combinations.

MCDA techniques and methods are useful to overcome these difficulties. The use of ordinal scale to score the different criteria does not allow the use of simple aggregation methods such as weighted sum. It is relatively easy to show the possible errors of the used weighted sum. Indeed, as FDA developed a sound tool for risk ranking (i-Risk) it will be easy to run this tool for a limited number of food-hazard combinations for which quantitative data are available, and then compare the ranking order obtained with i-Risk and the one obtained with the reported model.

To overcome the problem linked to the uncertainty about the scoring of the different criteria and the ordinal scale problem other MCDA techniques need to be deployed. One promising method is ELECTRE III.

II. RESPONSE TO CHARGE QUESTIONS

- 1. *The draft risk ranking model uses the FSMA statutory factors to define seven criteria for scoring food-hazard pairs.***
 - a. *Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.***

Multi-criteria decision analysis/approach (MCDA) in general follows the sequence below:

- Identifying objectives
- Identifying options/alternatives for achieving the objectives
- Identifying the criteria to be used to compare the options
- Analysis of the options
- Making choices, and
- Feedback

To adapt the MCDA to risk ranking we can define the sequence as following:

- Identify the objective: health impact that can be directly linked with food consumption (overall risk)
- Identify the list of food-hazard combination that contribute to the overall risk
- Identifying the criteria to be used to compare the contribution of the different food-hazard combinations to the overall risk
- Collect data for each food-hazard combinations relative to the identified criteria
- Aggregate the different criteria and rank the food-hazard combinations
- Continuous reassessment of the choices made in the past...

To answer to the question “*Are the seven criteria used in the draft model appropriate for a multicriteria decision approach?*” the objectives need to be clearly defined.

Multi-criteria analysis, in the current project, consists on the identification of food-hazard combinations that contribute most to the foodborne burden: risk ranking. The risk ranking has to be based on measurable criteria to assess the extent to which each combination contributes to the overall burden or risk.

The chosen criteria are:

Criterion 1: Frequency of outbreaks and occurrence of illnesses (i.e., Epidemiological Link)

Criterion 2: Severity of illness, taking into account illness duration, hospitalization and mortality

Criterion 3: Likelihood of contamination

Criterion 4: Growth potential/shelf life

Criterion 5: Manufacturing process contamination probability/intervention

Criterion 6: Consumption

Criterion 7: Economic impact

Criterion 1: This criterion is appropriate and relevant to distinguish between a high risk food-hazard high combination and low risk one in this ranking problem. However, this criterion is applicable only if data for all the food-hazard combinations are available and it has not the same interpretation for all the hazards: total number of cases including or not sporadic cases.

Considering the decision maker perspective, I think it is important to consider separately the two dimensions: 1) Total number of cases including outbreaks and sporadic cases, and 2) the number of outbreaks. For two food-hazard combinations with the same total number of cases, the one with outbreaks may be considered at highest risk. In the current report sporadic cases are considered in C7 scoring and are not included in C1 scoring. This choice is not logical, in my opinion, because it is done only for hazards for which outbreaks are observed or expected. My proposal is not to combine the two sub-criteria frequency of outbreaks and occurrence of illnesses and have two separate criteria: total number of cases and frequency of outbreaks.

Criterion 2: Again as in criterion 1, the interpretation of criterion 2 is not measuring the same things for all the hazards. The hospitalization rate and mortality rate can be used for both chronic and acute exposure. For example, if a chronic exposure to a carcinogenic chemical substance the known type of cancer will inform on the rate of hospitalization (almost 100%) and mortality rate.

Criterion 3: Likelihood of contamination. In general, in MCDA, the criteria are independent.

Criterion 3 is directly linked to criterion 1. The more likelihood of food-hazard combination contamination is, the higher frequency of illness occurrence might be observed. Criterion 1 is a result of exposure to a particular hazard through the consumption of a particular food. The

correlation between these two criteria needs to be addressed when combining the different criteria.

Criterion 4: Growth potential/shelf life. Same comment as for criterion 3. There is a possible correlation between criterion 4 and criterion 1.

Criterion 5: Manufacturing process contamination probability/intervention. This criterion is in principle correlated to criterion 3. When looking to the attributed scores, it was surprising to find food-hazard combinations scored 0 for criterion 3 being scored 5 for criterion 9 (11 food-hazard combinations out of 1286 food-hazard).

Criterion 6: Consumption. The percentage of population consuming food will not capture the entire consumption pattern. Information about frequency of consumption may contribute to the final risk. It will be interesting to know if the food is consumed daily, weekly, monthly, etc.

Criterion 7: Economic impact. This criterion should be applied to each hazard and not to each food-hazard combination. To avoid counting two times the total number of cases, it would be better expressed as the average economic impact per case of illness.

- b. ***Within the bounds of the FSMA-mandated factors, are there additional criteria beyond the seven criteria that should be considered? If so, please describe these additional criteria that might be considered and why.***

Figure 2-1 is an effort to explain the relationship between the seven criteria and the FSMA-mandated factors; however factor (v) is not covered. My interpretation of (v), “the likelihood that consuming a particular food will result in a foodborne illness due to contamination of the food”, is different from the one proposed in the current report. In my opinion, factor (v) is not only about exposure but also includes the dose-response relationship. That is, factor (v) tries to capture the infectivity or the toxicity of the hazard. In addition, factor (ii) is not totally covered because criterion 2 is not considering the level of contamination for chemical hazards, unless the frequency of contamination is taking into account only events with high level of contamination.

2. Are the scoring definitions for all criteria appropriate?

I will say that the definitions are in general clear and the provided document allows for the needed verifications. Please see 2b.

- a. ***Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and undeclared allergens)? If not, please describe changes that might be considered and why.***

No comment.

- b. Is the value function 0-1-3-9 and scoring matrix appropriate for the intended purpose to inform the designation of high-risk foods? If not, please describe changes that might be considered and why.***

Observed data and information related to the different criteria are grouped into scoring bins, which are defined and assigned a numerical value from 0 to 9. The numerical values from 0 to 9 reflect categories and their assignment is arbitrary, except that the values reflect the increasing quantity of the criteria being measured. The scale of the scoring is ordinal despite the distance between the possible values 0-1-3-9. The scoring 0-1-3-9 cannot be interpreted as an interval scale or as a ratio scale. Adding ordinal scales cannot be done in principle. Adding ordinal scales may lead to incorrect conclusions and this is the main challenge for risk ranking based on ordinal scales and the motivation for developing adequate combination of criteria measured with ordinal scales. The proposed ordinal scales intentionally leave gaps between the numerical values to better represent the assumed distance between categories is a possible approach but does not solve the entire problem.

The gaps between the numerical values are not consistent for all the criteria. For example, the significance of zero is for criteria absence of event and for other absence of data. The corresponding bins for 1-3-9 for one criteria are defined using a linear scale (i.e., consumption: 0 (>1%), 1 (1-5%), 3 (5-10%), 9(>10%)) and for another a sort of log-scale (i.e., contamination: 0 (No known occurrence), 1(<0.1%), 3(0.1-1%), 9(>1%)).

- 3. Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.***

Aggregation

Considering the nature of the measurement scales, I think that the proposed algorithm that combines criteria scores and weights is not appropriate.

Before presenting the suggestions to improve the current algorithm, it is important to know exactly the significance of the overall score. Because of the possible non-independency between the seven criteria, there is a need to create a sort of hierarchy between the seven criteria. One of the possible restructurings is as follow:

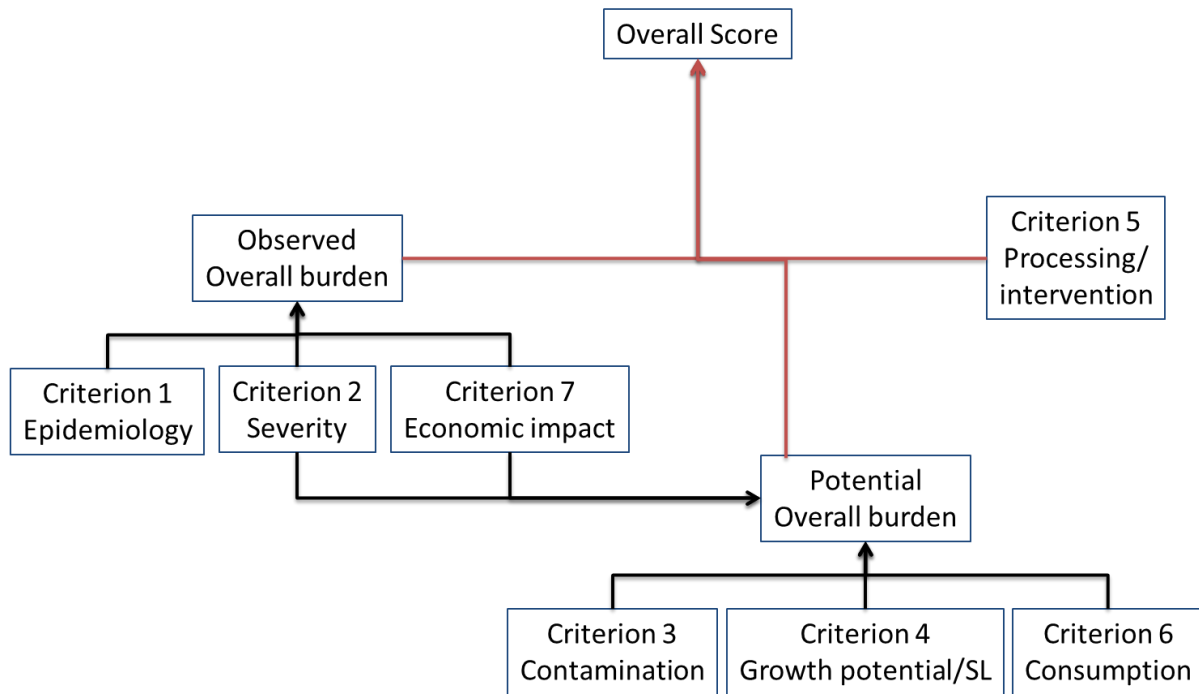


Figure 1: Risk ranking structure. Criterion 1 includes sporadic cases, criterion 7 economic impact per case of illness.

The assumptions of the new structure are:

- A food-hazard combination is assumed to be at high risk if the observed overall burden is high, or the potential overall burden is high or the criterion 5 (probability of contamination at processing is high and weak intervention).
- The observed overall burden is an assessment of the risk based on empirical epidemiological evidence: “top down” assessment
- The potential overall burden is an assessment of the risk based on empirical food chain evidence (contamination, consumption, growth...): “bottom up” assessment.

Two rankings may be performed: one considering “Observed overall burden” and the other “Potential overall burden”.

Using the report structuring of the criteria or the one proposed in Figure 1, the combination of the score attributed to the different criteria need to be modified:

As the scores are assigned using ordinal scales, it is not possible to use the weighted sum to aggregate the seven criteria. One of the possible solutions is to use the outranking concepts (Roy, 1978). An outranking relation of two food-hazard combinations a and b, as a binary relation S defined on a set of food-hazard combinations A, such that aSb (a outranks b) if, given what is known about the decision maker’s preferences, and given the evaluations on food-hazard combinations and the nature of the problem, there exist enough arguments to decide that a is at least as risky as b, while there is no essential reason to disapprove that statement. To implement the outranking concept, one can use one of the ELECTRE methods (Elimination and Choice Translating Reality, ELECTRE I, II, III, IV). ELECTRE III method was designed to deal with inaccurate, imprecise or uncertain data. This method utilizes pseudo-criteria instead of the defined and is suitable if at least one of the following situations is shown:

- Ranking problem where the ranked objects are evaluated (for at least one criterion) on an ordinal scale or on a weak interval scale. These scales are not suitable for the comparison of differences.
- A strong heterogeneity related to the ways criteria are evaluated which makes it difficult to aggregate all the criteria in a unique and a common scale
- Compensation of the loss on a given criterion by a gain on another one may not be acceptable for the decision maker. Then, such situations require the use of no compensatory aggregation procedures.
- For at least one criterion small differences are not significant in terms of preferences, while the accumulation of several small differences may become significant. This requires the introduction of discrimination thresholds.

The main purpose of ELECTRE III method is to rank the food-hazard combinations based on two indices, the concordance index and the discordance index defined for each pair of food-hazard combinations a and b.

The concordance index $c(a,b)$ is calculated by:

$$c(a,b) = \frac{1}{W} \sum_j w_j c_j(a,b)$$

where j is one of the seven criteria, W is the sum of weights of the different criteria (w_j) and

$$c_j(a,b) = \begin{cases} 1 & \text{if } sj(a) - sj(b) < q_j \\ 0 & \text{if } q_j < sj(a) - sj(b) < p_j \\ \frac{p_j - (sj(a) - sj(b))}{p_j - q_j} & \text{if } sj(a) - sj(b) > p_j \end{cases}$$

$sj(a)$ is the score for criterion j for food-hazard combination a, p_i and q_j are discrimination thresholds that define zones of strict difference, indifference and weak difference.

The concordance index evaluate to what extent (a) is at least as high risk as (b).

$$d_j(a,b) = \begin{cases} 1 & \text{if } sj(a) - sj(b) > v_j \\ 0 & \text{if } p_j < sj(a) - sj(b) < v_j \\ \frac{p_j - (sj(a) - sj(b))}{p_j - q_j} & \text{if } sj(a) - sj(b) < p_j \end{cases}$$

Where v_j is the veto threshold.

The overall concordance and discordance indices are then combined to obtain a valued outranking relation with credibility that a outranks b defined as:

$$\delta(a,b) = \begin{cases} c(a,b) & \text{if all } dj(a,b) \leq c(a,b) \\ c(a,b) \prod_{i \in J(a,b)} \frac{1 - dj(a,b)}{1 - c(a,b)} & \end{cases}$$

Where $J(a,b)$ is the set of criteria j for which $dj(a,b) > c(a,b)$.

The discordance Index to what extent the overall difference between the scores of (a) and (b) is enough important that (a) is not as high risk as (b).

The overall concordance and discordance indices are then used to provide two rankings: descending and ascending ranking. And the combination of the two ranking provides the final ranking.

The implementation of ELECTRE methods can be done using available software (<http://www.lamsade.dauphine.fr/english/software.html>).

References ELECTRE III:

- Rainer Bruggemann¹ and Lars Carlsen. Incomparable - What Now, III. Incomparabilities, Elucidated by a Simple Version of ELECTRE III and a Fuzzy Partial Order Approach. MATCH Commun. Math. Comput. Chem. 73 (2015) 277-302
- José Figueira, Salvatore Greco, Bernard Roy and Roman Slowinski. Electre Methods: Main Features and Recent Developments. HAL Id: hal-00876980 <https://hal.archives-ouvertes.fr/hal-00876980> 25 Oct 2013.

Weights

The approach used to elicit the possible weight is not well formalized.

First proposal: organize an expert elicitation using for example AHP method.

Second proposal: use constraint algorithm (sum of weight equal to 1) and learning system. The learning system can be organized by providing decision makers/stakeholders a random set of valuated criteria and ask the participants to rank them without providing any explanation about their outcomes.

- 4. *Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided?***
 - a. *Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.***

See my comment in 3.

- 5. *Considering the various scenarios (described in Section 7.3 and Appendix N) to aggregate food-hazard pairs in order to identify the foods which should be identified as high-risk vs. not high-risk, which option(s) are more appropriate to consider and why?***

It is hard to understand all the rationale behind each tested approach.

Alternative 1: not acceptable, for the same reason I explained previously for the aggregation of the seven criteria.

Alternative 2: the compensation is not acceptable. High score with one hazard will be compensated by low score with another hazard. Can we say that one food is on average safe or on average at high risk?

Alternative 3: it is one possible option to reduce the number of hazards per food.

Alternative 4: this option will underestimate the role of food with several hazards. Not acceptable

Alternative 5: Not acceptable. Because we are mixing factors that have not the same impact in regard to the different hazards.

a. Are there additional aggregation method(s) that might be considered? Please explain.

Another option is to use ELECTRE III to rank the food. The criteria will be the ranks obtained for each hazard. The pairwise comparison will allow calculating concordance and discordance indices based on the different ranks obtained for each hazard. The discrimination thresholds will be the distance between ranks. Then we can define strict difference if for example the difference between ranks is higher than 10. When a hazard is not considered for one food, the advantage will be given to the other food. However to avoid counting discordances corresponding to hazard ranked down a threshold may be chosen (i.e., <100 to be included as advantage if the food is not concerned by a given hazard).

Example of pairwise comparison: Just an example without taking into account for discrimination thresholds.

Hazard	Dairy - Ice Cream and Related		Seafood - N.E.C.	
	Advantage	Rank	Advantage	Rank
<i>Campylobacter</i> spp.	0		1	37
Ciguatoxin	0		1	55
<i>Clostridium botulinum</i>	0		1	75
Hepatitis A virus	1	19	0	20
<i>Listeria monocytogenes</i>	1	13	0	95
Norovirus	1	28	0	69
<i>Salmonella</i> spp.	1	15	0	44
Scombroid toxin (Histamine)	0		1	55
<i>Shigella</i> spp.	0		1	32
<i>Staphylococcus aureus</i>	1	25	0	35
STEC O157	1	14	0	114
Undeclared allergens	0		1	55
Undeclared allergens (other than fish)	1	5	0	20
<i>Vibrio parahaemolyticus</i>	0		1	

If you are interested by the idea, I can provide more details.

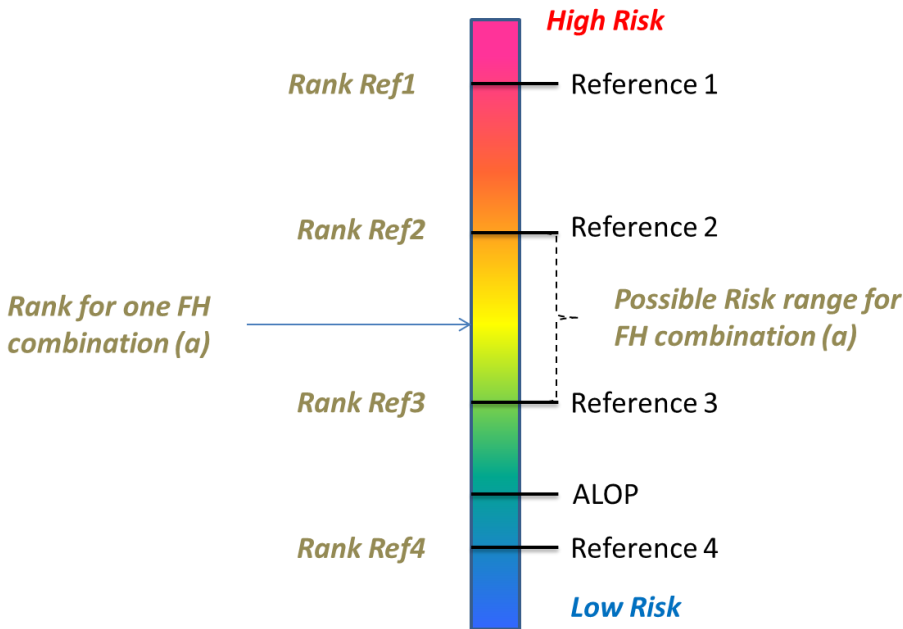
6. Given the underlying data supporting the scoring, what are the considerations to take into account when identifying high risk vs. not high risk food-hazard pairs or foods?

The system needs to be calibrated using external data. See my next comment.

a. What are the pros and cons in establishing a threshold considering all three types of hazards in the model, vs. drawing a line separately for microbial, chemical and undeclared allergens in foods?

A threshold without external data is actually difficult. One option is to consider a set of food-hazard combinations for which quantitative data and quantitative risk assessment are available

and use them to create a set of references to be compared with the U.S. ALOP (acceptable level of protection, i.e., 1/million DALYS per year per consumer)... it will be a way to have a sort of risk graduations:



7. Are the seven criteria and scoring definitions implemented appropriately in the SAS codes (Appendix P) and the Access Model?

The codes are correct. However, there are some differences in the score in the Excel files and the final score in Access tables.

Just one example of differences:

19	Fluid white milk, Grade A, pasteurized	88	<i>Campylobacter</i> spp.	5	1	9	1	4	2.	201	186
35	Fresh herbs - Group	231	<i>Cyclospora cayetanensis</i>	1	3	9	1	5	3.	201	159
				2					8	3	3

- a. Does the scoring logic described in Section 4 Figures 4-1 to 4-7) appropriately represent the scoring definitions described for each of the criteria in Section 3? If not, please describe what changes need to be made to correct it.

Yes.

- b. Are the scoring logic and order of preference accurately implemented in the SAS codes? (Please select 2-3 out of the 7 criteria for this evaluation). If not, please specify what changes need to be made.

Yes.

- c. Are equations 1 through 4 and data weighting factors accurately implemented in the model (either the SAS codes or the Access Model)?***

Yes.

- 8. In the Access Model, is the underlying relational database including lookup tables and algorithm appropriately designed and implemented? If not, please explain what changes should be considered.***

No comments.

- 9. Is the user interface of the Access Model sufficiently described for the user to understand each component of the model, e.g., foods, hazards, ranking criteria, results, and cited references?***

Yes. But I get an error message when I run the ranking. But when I update the outputs I get the results. When exporting the results to Excel, all the information about the scenario is exported but not the scores and ranking outputs. So I just copy and paste from the window interface.

- 10. How often should the model be updated, considering the data sources and data currently available and types of data that might become available in the future?***

To answer this question, I need to know the robustness of the ranking. This can be done by moving for each food-hazard combinations the scores and see when the rank is significantly different. For the non-robust ranks it is needed to look at uncertainty scores. If the uncertainty score is high and the rank is not robust that means the score of those food-hazard combinations needs to be updated as soon as possible.

- 11. Is the draft report clear in its description of the risk ranking approach, criteria and scoring definitions, and model limitations? If not, please identify which aspects are unclear or could be more transparent.***

In general, the report provides the needed information to understand the risk ranking process.

- 12. Do you have any additional comments? Please share them in your review.***

The sensitivity analysis should not be restricted only to the choices of criteria weights. A very simple sensitivity analysis can be made by ranking the food-hazard combination with and without each of the seven criteria and see if there are some significant changes in the rank order of food-hazard combinations.

As an example Figure 2 shows the changes of rank order when C1 is not included.

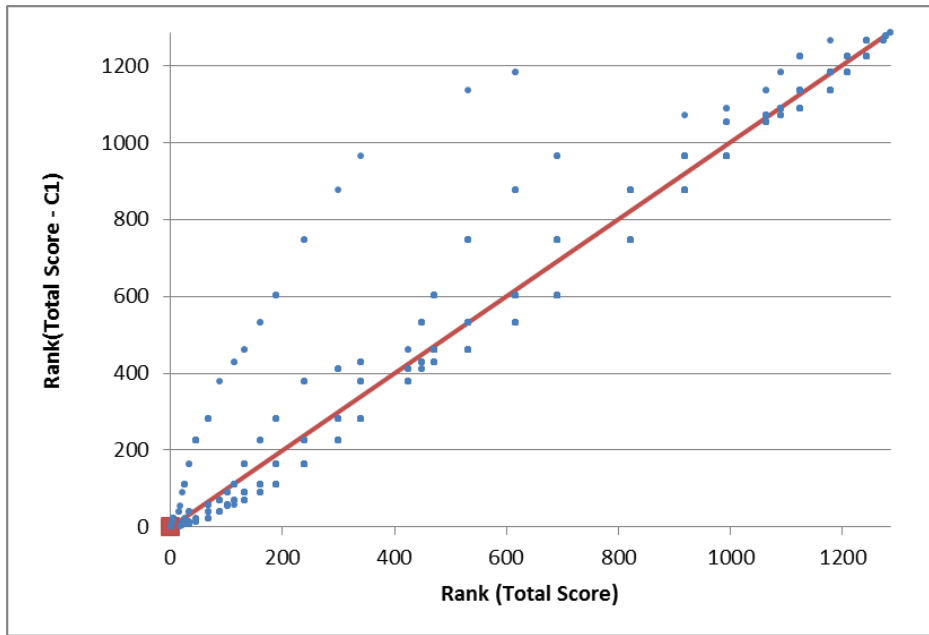


Figure 2. Impact of C1

To evaluate the overall importance of factor C1 on the risk ranking outputs, a statistic such as The Kendall’s tau rank correlation coefficient can be used.

Moreover for each food-hazard combination it will be useful to determine if a single score can modify significantly the risk rank order: this can be done by changing the score by zero only for the considering food-hazard combination and calculating the distance between the rank with all the scores and the one obtained when one score is excluded.

III. SPECIFIC OBSERVATIONS ON DRAFT REPORT FOR PEER REVIEW: RISK RANKING MODEL FOR PRODUCT TRACING AS REQUIRED BY SECTION 204 OF FSMA (RRM-PT Draft Report) WITHIN THE CONTEXT OF THE MODEL ITSELF.

Page	Paragraph/Line	Comment
		No comments

IV. SPECIFIC OBSERVATIONS ON APPENDICES TO THE DRAFT REPORT FOR PEER REVIEW: RISK RANKING MODEL FOR PRODUCT TRACING AS REQUIRED BY SECTION 204 OF FSMA (RRM-PT Draft Report) WITHIN THE CONTEXT OF THE MODEL ITSELF.

Appendix	Page/Row	Paragraph/Line/Column	Comment
			No comments

V. SPECIFIC OBSERVATIONS ON RISK RANKING MODEL FOR PRODUCT TRACING: USER’S GUIDE SHORT VERSION.

Page	Paragraph/Line	Comment
		No comments

VI. SPECIFIC OBSERVATIONS – PROVIDE SPECIFIC OBSERVATIONS, CORRECTIONS, OR COMMENTS ON THE ACCESS DATABASE *FDA'S HIGH RISK FOODS (HRF) MODEL*.

Menu Choice	Tab	Steps taken within the tab	Comment
			No comments

Reviewer #3

Peer Review Comments on FDA's Draft Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA:

HRF Model, the HRF Model Report, the HRF Model User's Guide, and the HRF Model Design Report

Reviewer #3

I. GENERAL IMPRESSIONS

Mitigating outbreaks of foodborne illness requires a consideration of the various risks associated with many different foods and food byproducts. The Food and Drug Administration (FDA) has developed a semi-quantitative risk ranking model that integrates several criteria and metrics associated with food assessment. The document offers a robust approach for the purpose of risk ranking that utilizes a multi-criteria scoring model to derive a risk score and ultimately inform decisions on food-based health risks.

Within this framework, the general idea is that foods with more elevated risk scores should be a higher priority for consideration. The use of multi-criteria models in risk prioritization are particularly suitable when it is impractical to build and populate a full causal risk model, so where multi-criteria models instead act as proxies that are measurable and which are thought to be associated with risk. The proposed approach integrates significant volume of historic and measurable data with expert judgment.

With this said, I have some concerns about specific algorithms used. It seems the risk factors should be multiplicative because they form a chain of events leading to consequences, but the document utilizes additive model. Additive model may be a reasonable approximation, but additional clarification and justification may be required. In any event, there are no easy shortcuts to producing a score that has the precision of a structured risk analysis and the proposed approach is a major improvement over traditionally used dashboards or individual risk indicators.

II. RESPONSE TO CHARGE QUESTIONS

1. *The draft risk ranking model uses the FSMA statutory factors to define seven criteria for scoring food-hazard pairs.*
 - a. *Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.*
 - b. *Within the bounds of the FSMA-mandated factors, are there additional criteria beyond the seven criteria that should be considered? If so, please describe these additional criteria that might be considered and why.*

The seven criteria selected by FDA seem to be reasonable and well-justified based on the body of available scientific knowledge, but detailed knowledge of FSMA is outside of my area of expertise.

2. *Are the scoring definitions for all criteria appropriate?*
 - a. *Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and*

undeclared allergens)? If not, please describe changes that might be considered and why.

Various types of hazards appear to be defined well, but detailed hazard identification for foods is outside of my area of expertise.

b. Is the value function 0-1-3-9 and scoring matrix appropriate for the intended purpose to inform the designation of high-risk foods? If not, please describe changes that might be considered and why.

It is unclear whether the proposed value function and scoring matrix are appropriate to inform the designation of high-risk foods. On one hand, it is certainly true that for particular assignments of scores, a higher score is associated with a higher risk. On the other, it is unclear whether the scores are meant to indicate the magnitude of a risk factor, its order-of-magnitude of a risk, or something else entirely. For example, is a 9 actually supposed to be three times as bad as a 3? In the example on p. 17, a hazard with a hospitalization rate of 9% would have a score of 1, and a hazard with a hospitalization rate of 21% would have a score of 9.

This only becomes more complicated and unclear for the case of non-numerical constructed scales. In a sense, the point of these scales is to serve as proxies for probabilities whose explicit elicitation will be impractical. It would help if the scales were constructed with consistent intent, e.g., with the intent that scores correlate with the magnitude of the risk. As will be discussed in my later comments, because these scores are to be used in further calculations, it is important to be clear on what they are supposed to represent, since their properties determine when it is appropriate to perform such calculations and how to interpret their results. For example, statistics courses teach that averages of ordinal data are not meaningful – are three satisfied customers the same as one dissatisfied customer, one satisfied customer and one extremely satisfied customer?

3. Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.

The overall approach is reasonable. In a perfect world, an ideal model (such as a full, vetted, detailed risk assessment model) would calculate explicit risks associated with different foods and compare them to benchmark levels derived using solid toxicology models. However, with various foods to consider and limited information to analyze in this vein, such a model is practically difficult to adequately develop. At the other extreme, a simple model might just require a food expert or panel of experts to give a qualitative score to each food and rank them accordingly. This proposed model lies in between, providing some structured knowledge base: the data fields used, the algorithm for calculating risk scores, and the weighting parameters and scoring protocols used are intended to get more of the precision and transparency of the ideal model without creating onerous information requirements. To the extent that it does this in a way that approximates what an ideal model would do, the approach can lead to real practical improvement in food prioritization by taking the best course of action to overcome current knowledge limitations.

With this said, I have concerns about specific algorithms used. It seems the factors should be multiplicative because they form a chain of events leading to consequences. Ideally, the overall

score would be highest for the risks with the greatest average consequences, i.e., foods that are classified as high-risk would have higher risk than foods that are classified as low risk. This may not happen if the hazard scores represent magnitudes, or are approximately linear in probability. For example, hazard A with a 50% chance of causing 5 units of harm has an expected loss of 2.5 units, while hazard B with a 90% chance of causing 2 units of harm has an expected loss of 1.8 units. However, if the factor scores are translated to (5, 5) for hazard A and (9, 2) for hazard B, hazard A has a lower aggregate risk score (10) than hazard B (11).

If created and applied with care, factor scores that reflect order-of-magnitude of risk (and that also represent sequential events) can be added in such a way that the higher score is associated with higher risk. This is due to the fact that adding such factor scores would be equivalent to multiplying the magnitude of risks at each stage. For aggregation within a food group, a mathematically sound approach along these lines (that proxies for risk analysis) is even harder to figure out. The problem is, there are no easy shortcuts to producing a score that has the precision of a structured risk analysis. This is better than nothing if the scores are handled as coarse indicators. Their numerical nature should not be confused with across-the-board rigor.

More details are presented in the additional comments at the end of this review.

- 4. *Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.***

Related to this subject, decision analysis centers on the idea that by quantifying the preferences for each *criteria* rather than specific food, a more objective and systematic prioritization framework can be achieved. By defining which criteria are most important, and integrating these respective weights with scores representative of foods performance by each criteria, an integrated risk score can be quantified. There are different ways to elicit weights and to assign scores, some of which are tailored to specific MCDA methodologies. In general, if the individual hazard scores represent order-of-magnitude risk factors, equal weighting is most appropriate from a mathematical perspective. If the individual scores represent additive indicators of some sort, then different weights could be used, where they would provide flexible scaling factors.

- 5. *Considering the various scenarios (described in Section 7.3 and Appendix N) to aggregate food-hazard pairs in order to identify the foods which should be identified as high-risk vs. not high-risk, which option(s) are more appropriate to consider and why? Are there additional aggregation method(s) that might be considered? Please explain.***

Depending on what the food-hazard pair scores represent, summing them to identify high-risk foods may or may not be an appropriate step. If they represent order-of-magnitude risks, then summing food-hazard pair scores would be inappropriate and a more detailed aggregation rule would be necessary. If they represent magnitudes of risks, then summing them, with some weighting corresponding to relative quantities consumed of the different foods, would be the appropriate course of action.

- 6. Given the underlying data supporting the scoring, what are the considerations to take into account when identifying high risk vs. not high risk food-hazard pairs or foods? What are the pros and cons in establishing a threshold considering all three types of hazards in the model, vs. drawing a line separately for microbial, chemical and undeclared allergens in foods?**

Perhaps there could be two different thresholds in use, including (i) one separated at the level of hazard type, and (ii) one collectively for the food. For example, either a collective score above 300, or a score on any of the three separate hazards above 200 might qualify as a high risk that merits special handling. The rationale for the former is what seems to underlie the whole approach, simply as an indicator of the overall danger. The rationale for the latter might be that emergent conditions related to the offending hazard could inflate the hazard to more worrisome levels, i.e., the food's risk estimate might have “fatter tails” if it is driven by one of the hazards.

- 7. Are the seven criteria and scoring definitions implemented appropriately in the SAS codes (Appendix P) and the Access Model?**
- a. Does the scoring logic described in Section 4 Figures 4-1 to 4-7) appropriately represent the scoring definitions described for each of the criteria in Section 3? If not, please describe what changes need to be made to correct it.**
 - b. Are the scoring logic and order of preference accurately implemented in the SAS codes? (Please select 2-3 out of the 7 criteria for this evaluation). If not, please specify what changes need to be made.**
 - c. Are equations 1 through 4 and data weighting factors accurately implemented in the model (either the SAS codes or the Access Model)?**

This reviewer cannot provide details on SAS codes and Access model since he has experience in risk assessment and decision analysis and does not have advanced knowledge in SAS and Access.

- 8. In the Access Model, is the underlying relational database including lookup tables and algorithm appropriately designed and implemented? If not, please explain what changes should be considered.**

This reviewer cannot provide details on SAS codes and Access model since he has experience in risk assessment and decision analysis and does not have advanced knowledge in SAS and Access.

- 9. Is the user interface of the Access Model sufficiently described for the user to understand each component of the model, e.g., foods, hazards, ranking criteria, results, and cited references?**

This reviewer cannot provide details on SAS codes and Access model since he has experience in risk assessment and decision analysis and does not have advanced knowledge in SAS and Access.

- 10. How often should the model be updated, considering the data sources and data currently available and types of data that might become available in the future?**

[The reviewer did not comment.]

11. Is the draft report clear in its description of the risk ranking approach, criteria and scoring definitions, and model limitations? If not, please identify which aspects are unclear or could be more transparent.

The report is rather comprehensive and appears to be clearly written. It has an excellent and logical structure. Assumptions are clearly stated, and Appendices are very useful in evaluating data and calculation algorithms. The main concern is about more clarity on the meaning of the scoring rules.

12. Do you have any additional comments? Please share them in your review.

Additional clarification for Question 3.

The proposed approach would be something like this:

For food-pathogen pair i risks r_{ij} are defined verbally as relating to multiplicative factors such as probabilities of occurrence, magnitude of loss per given occurrence etc. so that total expected loss to a person eating a serving of the food is $r_i = r_{i1} * r_{i2} * \dots * r_{in}$. This approach uses an additive risk score and calculating total risk $x_i = x_{i1} + x_{i2} + \dots + x_{in}$. Ideally, this score works such that $x_1 > x_2$ if and only if $r_1 > r_2$ - i.e., you would never classify food A as high risk and B as low risk when the expected loss of A is lower than the expected loss of B. The scoring rules where x_{ij} are given scores from 1 to 9 will only work if x_i is proportional to $\ln(r_i)$. It is not at all clear this is the case. Visual inspection suggests that x_i have an approximately linear relationship with r_i . Even if x_i are appropriate, it does not make sense here to use weights, since if we take exponentials to convert x to r , coefficients on the x terms transform to powers on the r terms (each of which ought to be linear in expected loss).

When considering risk for a food across multiple food pathogen pairs, we cannot simply add their scores. Instead, we would have to calculate $x_{12} = \ln[\exp(x_1) + \exp(x_2)]$. Similarly, when considering risks for multiple similar foods (e.g., tuna and salmon), we cannot simply sum them. First, we want to weight by proportion - $r_{\text{fish}} = \text{proportion_tuna} * r_{\text{tuna}} + \text{proportion_salmon} * r_{\text{salmon}}$, or more generally, if food group i consists of subgroups k , $r_i = \text{sum over } k \text{ of } w_{ik} * r_{ik}$. To convert these back to total scores consistent with the kind of x -values I have described for multiplicative risk factors, we would need to have $x_i = \ln[\text{sum}(w_{ik} * r_{ik})]$, and since there is no neat form for $\ln(a+b)$, the whole thing becomes kind of unwieldy if done right.

In sum, this represents a potentially serious methodological problem that must be directly addressed. Although algorithms would not be difficult to modify, they would be less transparent for non-expert users.

The overall approach might be solidified by:

- (1) clarifying whether factor scores are meant to indicate magnitude or order-of-magnitude of probabilities/consequences
- (2) ideally they would represent order of magnitudes
- (3) the method would be used as-is to produce scores for the food-hazard pair

- (4) to calculate collective risk over hazards we would sum the exponential of the hazards and then take the logarithm of that total
- (5) to calculate risk for a food group we would calculate a weighted sum of the exponential of scores for each constituent food and take the logarithm of that total.

III. SPECIFIC OBSERVATIONS ON DRAFT REPORT FOR PEER REVIEW: RISK RANKING MODEL FOR PRODUCT TRACING AS REQUIRED BY SECTION 204 OF FSMA (RRM-PT Draft Report) WITHIN THE CONTEXT OF THE MODEL ITSELF.

Page	Paragraph/Line	Comment
		No comments

IV. SPECIFIC OBSERVATIONS ON APPENDICES TO THE DRAFT REPORT FOR PEER REVIEW: RISK RANKING MODEL FOR PRODUCT TRACING AS REQUIRED BY SECTION 204 OF FSMA (RRM-PT Draft Report) WITHIN THE CONTEXT OF THE MODEL ITSELF.

Appendix	Page/Row	Paragraph/Line/Column	Comment
			No comments

V. SPECIFIC OBSERVATIONS ON RISK RANKING MODEL FOR PRODUCT TRACING: USER'S GUIDE SHORT VERSION.

Page	Paragraph/Line	Comment
		No comments

VI. SPECIFIC OBSERVATIONS – PROVIDE SPECIFIC OBSERVATIONS, CORRECTIONS, OR COMMENTS ON THE ACCESS DATABASE FDA'S HIGH RISK FOODS (HRF) MODEL.

Menu Choice	Tab	Steps taken within the tab	Comment
			No comments

Reviewer #4

Peer Review Comments on FDA's *Draft Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA:*

HRF Model, the HRF Model Report, the HRF Model User's Guide, and the HRF Model Design Report

Reviewer #4

I. GENERAL IMPRESSIONS

Detailed exploration of the draft model and its accompanying documentation suggests that there are anomalies in the estimates produced by the model compared to what would be expected from the available epidemiological data.

At least some of these anomalies seem to arise from logical errors in the model or that the model does not accurately represent the mathematical relationships between risk-affecting factors. For example, some of the problems seem to arise from inappropriate combinations of factors that scale on algebraic scales and those that naturally scale exponentially. In the consideration of methods for aggregation, one proposed approach cannot be justified logically. Also, all aggregation approaches introduce a logical error based on erroneous amplification of consumption scores, and there seems to be an anomaly in the implementation of the weighting factors as they apply to uncertainty and confidence scores.

In general, the apparent problems in the model seem to arise from trying to use a semi-quantitative risk matrix approach. The simplification of risk-affecting factors that are described by continuous variables into sets of three (or four) categorical values, introduces demonstrable anomalies in the scoring and distorts the relative risk estimates. Given that the RRM is developed in computer software that is capable of complex computations, and that there is no apparent need to simplify the calculations, e.g., if the model needed to be able to be used by people without access to a computer, the reasons for this approach are unclear. A better outcome would be expected by preserving the maximum value in the (quantitative) data and building a risk estimation model that achieves that.

In places the discussion in the report seems more about the mechanics of the modelling process and interface function, rather than in the underlying logic that leads to model predictions and evaluation of their credibility.

The report itself contains many minor presentation errors.

In short, the errors apparent in the model and in the accompanying documentation suggest that the model is not yet ready for its intended use.

II. RESPONSE TO CHARGE QUESTIONS

1. *The draft risk ranking model uses the FSMA statutory factors to define seven criteria for scoring food-hazard pairs.*

a. *Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.*

- b. ***Within the bounds of the FSMA-mandated factors, are there additional criteria beyond the seven criteria that should be considered? If so, please describe these additional criteria that might be considered and why.***

Food safety risk is normally interpreted as the number and severity of *potential* or actual cases of foodborne illness related to some defined food(s) and hazard(s). To assess such risk its necessary to consider the probability that a food will be contaminated with the hazard of interest, the number of people who eat that food and how frequently, the probability that the level of contamination will be high enough to cause illness, and severity of the resultant illness. “Severity” encompasses the severity of disease symptoms and duration of illness experienced by a consumer (often evaluated using the DALY or QALY metrics) but may also consider the *absolute* number of cases. For microbial hazards, the probability of contamination at a level likely to cause illness is also related to the potential for growth of the organism in the food. The potential for growth is a function of the specific ecophysiology of the hazardous organism, the composition and packaging of the food (esp. gaseous atmosphere) and the time, temperature, (and to a lesser extent) the relative humidity of the storage environment. In the Risk Ranking Model economic ‘impact’ was also required to be included as part of the overall risk, though the economic impact assessment was limited to health care costs, and did not include costs to industry from an outbreak, etc.

The factors nominated by FMSA for inclusion in the designation of HRF foods were:

- i) the known safety risks of a particular food, including the history and severity of foodborne illness outbreaks attributed to such food, taking into consideration foodborne illness data collected by the Centers for Disease Control and Prevention (CDC);
- ii) the likelihood that a particular food has a high potential risk for microbiological or chemical contamination or would support the growth of pathogenic microorganisms due to the nature of the food or the processes used to produce such food;
- iii) the point in the manufacturing process of the food where contamination is most likely to occur;
- iv) the likelihood of contamination and steps taken during the manufacturing process to reduce the possibility of contamination;
- v) the likelihood that consuming a particular food will result in a foodborne illness due to contamination of the food; and
- vi) the likely or known severity, including health and economic impacts, of a foodborne illness attributed to a particular food.

These nominated factors include consideration of factors generally considered to affect foodborne health risk, though in some cases in a somewhat cryptic way, e.g., the known safety risks from epidemiological data implicitly encompass many of the other factors that the more inductive approach to risk assessment considers explicitly. This potentially introduces a confounding in the approach to the risk ranking because some factors will, in effect, be considered twice. (This aspect was discussed in the complementary data review report).

In the above set of factors, however, the total consumption is not explicitly nominated as a risk-affecting factor, except that it is implicit in the epidemiological data because, all other things being equal, foods that are more frequently consumed are more likely to be involved in detectable outbreaks. This is not necessarily an oversight in the identification of risk-affecting

factors. In managing risk, or in identification of a “HRF”, it may be the *intrinsic* properties of a food and the processing, packaging and storage conditions that are considered to be the most important components for risk, i.e., to be able to estimate the *risk-per-serving*.

In the development of the risk ranking model the six factors above were translated into seven criteria:

- Criterion 1. Frequency of outbreaks and occurrence of illnesses (i.e., Epidemiological Link);
- Criterion 2. Severity of illness, taking into account illness duration, hospitalization and mortality
- Criterion 3. Likelihood of contamination
- Criterion 4. Growth potential/shelf life
- Criterion 5. Manufacturing process contamination probability/intervention
- Criterion 6. Consumption
- Criterion 7. Economic impact

The mapping of these seven criteria to FSMA’s six factors is shown diagrammatically in Figure 2-1 in the Draft Report, which clearly shows the some of the criteria are confounded, e.g., as suggested above, the epidemiological data would be expected to be a reflection of other risk-influencing factors (Criteria 3 – 6), so that effectively a “double-counting” of some factors is occurring in the logic of the model. Notably, in the FDA model, consumption is explicitly added to the risk ranking criteria, despite that it appears not to have been explicitly considered by FMSA. As such, there may need to be clarification of whether the intent of the RRM is to rank risk on a per-serving basis, or on a whole of production/supply basis.

In the draft model, however, one necessary risk-affecting factor that appears to be inadequately addressed is the relevance of the *level* of contamination, i.e., consideration of the dose-response relationship and its influence on the inferred probability of illness. This issue is implicit in FMSA factor v). It is also apparent from Figure 2-1 that it is *not* explicitly considered in the draft risk ranking model, i.e., because factor v) is ‘covered’ only by consumption data. The lack of an infectious/toxic dose consideration seems curious given that the draft model is stated to have been based on FDA’s fresh produce risk ranking tool which *does* consider ‘infectious dose’.

Another consideration that is, perhaps, implicitly suggested in FSMA criterion iii), but not explicitly considered in the FDA draft risk model, is the influence of the *sequence* of risk affecting influences. An obvious example is a heating step for a food in a sealed container, or a cooking step prior to consumption, that eliminates a microbial risk completely. Conversely, if recontamination and growth is possible after a ‘kill’ step, the risk will be completely different. The draft model does not reflect the significance of such differences in the *sequence* of steps and cannot represent this using the additive scoring process adopted, as will be illustrated more clearly in responses to later questions.

2. Are the scoring definitions for all criteria appropriate?

- a. Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and undeclared allergens)? If not, please describe changes that might be considered and why.**

The hazards are appropriately described for microbial, chemical and allergen hazards. Some of the chemical hazards listed are of microbial origin (algal and fungal toxins), but are usually already present in the food (prior to processing) and would not increase in level during processing, distribution etc. Histamine levels can increase as microbial metabolism continues, and this difference is appropriately identified in the approach developed, and dealt with by considering histamine as a microbial hazard. Nonetheless, fungal growth and aflatoxin production is also possible in many foods after processing, and the difference in approach for histamine, compared to aflatoxin, requires further explanation for transparency of the document and approach.

The relativities of some scores do not seem correct. For example, hazard severity scores for microbial hazards vs. allergens do not accord with independent expert assessment by Minor *et al.* (2015) and it was notable that 70% of the highest ranked hazards for all food-hazard pairs were for allergens, based only on expert opinion rather than data. This does not seem to accord with the available public health data. There is more detailed discussion of these anomalies in the complementary data review report.

- b. Is the value function 0-1-3-9 and scoring matrix appropriate for the intended purpose to inform the designation of high-risk foods? If not, please describe changes that might be considered and why.**

Additive scoring schemes, even if based on exponential scales, cannot capture the full complexity of factors and their interactions that contribute to food-borne risk.

All scoring schemes that attempt to translate continuous variables affecting risk into a limited number of categories and assign discrete numerical values will distort the calculations and relativities of risk. This compromise is sometimes necessary or useful when having to do many sets of calculations with imperfect data and for diverse properties. The sort of semi-quantitative scoring scheme adopted here is useful when one does not have access to a computer to do the complex calculations involved in quantitative risk assessment. However, given that the system described here is for government use, and is already presented as relatively sophisticated computer software, the reasons for adopting what is essentially a complex decision matrix rather than a generic quantitative risk assessment model, e.g., *iRisk*, are unclear and seem to require further explanation in the draft RRM report and accompanying documents and tools. The relationships between risk-influencing factors and how they contribute to risk can be expressed algebraically so that the full value of numerical data and the understanding of the interactions of factors that lead to risk could be preserved. It may not have been possible in this risk ranking project, e.g. due to time or personnel constraints, to collate and analyze all the data for each food-hazard pair into a representative value and, in those circumstances, a semi-quantitative approach as adopted may make sense. However, further justification/explanation of the adoption of a semi-quantitative risk matrix approach would be useful and contribute to the transparency of the approach and model.

Prima facie, the use of a quasi-exponential approach in scoring (i.e., where a difference of three in a score typically represents a factor of ten difference in risk, or a score difference of 9 represents a thousand-fold difference in risk) also makes sense to reflect that most (though not all) food safety risk-affecting risk factors multiply together to generate the risk. However, microbial risk assessments, in particular, often involve calculations with, and combination of, factors that operate naturally on exponential scales and others that operate naturally on arithmetic scales. This can lead to problems in semi-quantitative assessments based on discrete categories, as illustrated below.

Discretization of continuous variables into a small number of ‘representative’ values, or categories, can be expected to produce anomalies. Consider, for example, two product:hazard pairs. Let hazard ‘A’ be at the lower end of the ‘middle’ range for three factors and hazard ‘B’ be at the upper end of ‘middle’ range for the same three factors. Each range represents roughly a factor of 10. Thus, while both hazards would score the same (i.e., $3+3+3 = 9$), the true difference in risk would be expected to be closer to 1000-fold, i.e., hazard ‘B’ should have a risk score 9 higher than hazard ‘A’. As a second example, imagine that hazard ‘A’ is at the upper end of the ‘middle’ range for 3 factors, while hazard ‘B’ is at the lower end of the ‘high’ range for 3 factors. Now hazard ‘A’ has a score of 9, while hazard ‘B’ has a score of 27. On the basis of the quasi-exponential scale, the risk from hazard ‘B’ is now evaluated to be 1 billion times greater than the risk from hazard ‘A’ even though we know they are only subtly different in their levels in each of the three categories.

As a further example, consider three pairs of food:hazards. Hazard 1 has three attributes at the upper end of the ‘low’ range, and has a risk score of 3 ($1+1+1$). Hazard 2 has three attributes in the middle on of the ‘intermediate’ range and receives a score of 9 ($3+3+3$). Implicitly, Hazard 2 is 1000-fold more risky than Hazard 1, yet the score difference is 6 (implicitly only two factors of 10). Now, consider Hazard 3 which has three attributes at the low end of the relevant ‘high’ range. It would be scored 27 but, implicitly Hazard 3 is ~1000-fold more risky than Hazard 2. The score difference, however, is 18 suggesting 6 orders of magnitude difference, rather than the 3 orders ‘built into’ this scenario. The relative risk from Hazard 2 compared to Hazard 1 is the same as the relative risk from Hazard 3 to Hazard 2, yet the score differences are completely different. Thus, when used in scoring the overall risk, the relative risk scales inherent in the scoring for individual criteria become distorted, and the final risk score seems more like an absolute scale rather than an exponential scale. The combination and interconversion of factors that operate on arithmetic *cf.* logarithmic scales is a recognized source of logical errors in food safety risk assessment particularly for risks from microbial hazards.

As another example (mentioned earlier) of logical weakness in the scoring scheme, a listericidal process in a hermetically sealed product effectively eliminates the risk from that hazard in that product, irrespective of what occurred before or what will happen later to that product, as long as package integrity is preserved. Equally, proper cooking can completely eliminate many hazards. The current scoring scheme does not enable a “reset to zero” for ‘cidal’ processes and does not correctly represent such scenarios.

These examples are not presented to suggest that the model requires change, but that *users will have to be very careful about the interpretation of the risk ranking results*, particularly if one is

trying to assess relative risk, ***and particularly if trying to establish a threshold risk value that defines an HRF.***

Given the above discussion, however, there does appear to be a logical deficiency in the RRM and it is suggested that the approach/considerations used to define FDA's 'potentially hazardous foods' or 'temperature controlled for safety' foods (*see also* response to Q.6) may offer a good starting point, or useful insights, for the proposed risk ranking model despite that it will only offer insights about ranking of microbial hazards. It also would seem to offer the advantage of policy consistency.

As discussed more fully in the complementary report on the model data, while no simple scoring scheme is likely to correctly rank all product:hazard pairs and will of necessity be a compromise, it may be possible, nonetheless, to achieve consensus among stakeholders, i.e., to accept the RRM as being the best approach possible, or at least as good as any other approach, if stakeholders agree that the RRM does produce a ranking that 'make sense'. Users should remain aware, however, of those model limitations and potential anomalies and not accept the model results without some form of scrutiny, or 'reality checks'.

3. Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.

See comments above (in response to Q2b) which comments on and exemplifies potential problems in the scoring scheme rather than the "algorithm". An alternative approach was suggested there. The calculations, including selection of weighting factors are very simple and straightforward, i.e., simple additions of the seven criteria scores after application of any weighting. (In fact, the use of the term algorithm is not really relevant in this case because the calculation can be expressed as a single equation). Importantly the calculations are transparently presented and explained in the Draft Report with one possible exception (*see* response to Question 7b).

4. Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.

Before commencing the response to this specific question, it is worth noting that Figure 6-1 in the Draft Report presents a distribution for all product:hazard pairs. In the Draft Report it is suggested that useful information could be derived from the proportion of cases above, or below, certain "threshold" scores, e.g. that appear as inflection points on the plot. To place that notion in context, I developed a simple stochastic model using Analytica stochastic simulation software. The model had seven variables, each of which could have a value of 0, 1, 3 or 9. Equal probability (0.25) was assigned to each value for each of the seven variables. The model calculated the sum of the seven values. 1286 iterations of the model were executed, and the score for each iteration recorded. The scores were sorted, and then graphed, with the same axes as Figure 6-1 of the Draft Report. That plot is presented overleaf (Figure A) and was developed to

show that the shape of the plot is strongly influenced by the number of ways that intermediate scores can be achieved, compared to very high, or very low values. The similarity of the graph based on combinations of completely randomly derived values (Figure A), compared to the outputs of the model based on its ascribed values derived from data and expert opinion (Figure 6-1) suggests that caution needs to be exercised if using Figure 6-1 to derive “threshold” values for the designation of a HRF and, in particular, if intending to ascribe special meaning to inflection points on the plot as is hinted at in Section 6.2 of the RRM Draft report.

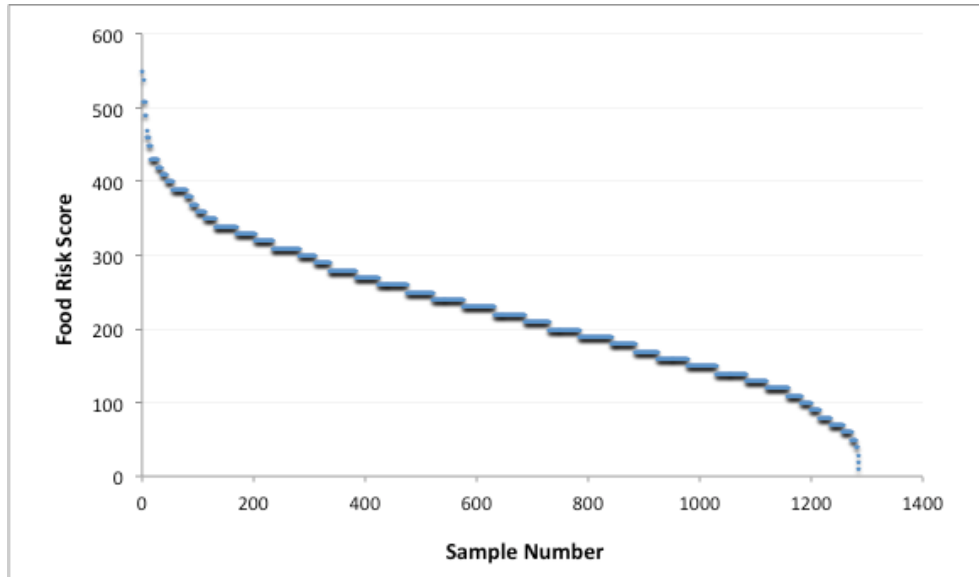


Figure A. Results of a simple stochastic model emulating the risk scoring system and showing the distribution of food risk scores that would arise from a *completely random* allocation of scores for each of seven criteria, illustrating the strong similarity to the scores distribution derived from all food:hazard pairs considered in the RRM.

At L.18 *ff.* on p. 85, it is noted that food:hazard pairs with a score of >310 have a majority of scores of 3 or 9 (in the absence of additional weighting, i.e., all weights equal to 10). This observation is a simple consequence of the scoring scheme itself. The additive seven-criterion scoring scheme is based on discrete score values of 0, 1, 3 or 9 for each of the seven variables. From that it is impossible to generate a score of 310 unless at least 3 of the scores are 9 and one of the scores is 3, i.e., the majority of the seven criterion scores. Thus, the discussion in this section of the Draft Report seems irrelevant and inane. The observations discussed in that section are about the consequences of the scoring scheme, not the underlying processes and phenomena that dictate the food safety risk, and which the scoring scheme/risk ranking model should aim to reflect.

Criteria weighting

Having conducted and been involved in several fully quantitative risk assessments, I can attest that I have no intuitive sense of which risk-affecting factors would be most important in a risk-ranking process, and should therefore be ascribed additional weight. Importantly, the factors included explicitly in a fully quantitative risk assessment do not readily correlate with the 'compound' variables included in the Draft Risk Ranking model under review. Similarly, the influence of a factor can vary according to the extent of variability defined for that factor and, if the variability in a factor changed significantly (e.g., the storage time for a product), that factor's influence on the risk estimate might also change and require more (or less) weight to be assigned to it in a semi-quantitative risk assessment scheme such as presented in the Draft Report. The issue only arises, however, if a scoring scheme, rather than a fully quantitative risk assessment model is used for generating the risk estimates.

I'd expect that different stakeholders would place more weight on certain kinds of information that they were more familiar with and that provided useful insights to them and their risk assessment, e.g., a food processor might weight product formulation/packaging and the rigor of the manufacturer's HACCP system more heavily than epidemiological data, because outbreaks only occur when there are system failures. Conversely, an epidemiologist might rely on disease statistics more heavily because they have less understanding of processing and distribution factors that might have affected risk, instead relying more on the types and quantities of food eaten and the relative susceptibility of the victims. Neither approach is inherently better or worse. Any approach based on stakeholder opinion/experience would need to select a wide range of expert stakeholders and also seek to develop a consensus approach. The consensus process might also need to consider the reliability of the various data sources used to provide estimates for the factors/criteria in the Draft Model.

As the discussion in Appendix M1 of the draft RRM report illustrated, there is no simple answer to the question of weighting, nor any single approach that is widely endorsed, despite that several apparently rigorous approaches have been articulated.

As stated earlier, as a risk modeler my preferred solution would be to use a model that explicitly considers all the fundamental factors that affect food safety risks, and to unambiguously describe the interrelationships between those factors, to develop rigorous and transparent risk estimates and rankings. Doing so would obviate the need for weighting factors.

Failing that, another approach implicitly suggested in Appendix M1 would be to conduct sensitivity analyses on the model to better understand which criteria had most influence. A problem with that approach is that the influence that those criteria already have may be inappropriate.

Another approach would be to conduct sensitivity analyses on existing fully quantitative risk assessment models to see which factors, or combinations of factors relevant to criteria used in the Draft Risk Ranking model have the most influence in the risk estimates generated and to derive appropriate weighting factors from that analysis.

Having experimented with the RRM model (i.e., via the Access model interface) and thought about relative weightings for Criteria and specifically the reliability of epidemiological

information and illness severity information - as opposed to the inferences about risk generated by the RRM - I began to wonder whether the “epidemiological link” Criterion and “severity” Criterion should be weighted MUCH more heavily than Criteria 3 to 6 which aim to infer the risk (essentially the same information as the epidemiological data), rather than relying solely on the (imperfect) epidemiological data. Thus criterion 1 might receive 30% weight, Criterion 2 would receive 20%, and Criteria 3 to 6 would receive 7.5% of the weight each, and Criterion 7 would receive 20%. In this manner the actual epidemiological data indicating frequency of illness would have the same weight as the inferred level of illness (inferred from Criteria 3 – 6) and both would be combined equally with the severity criteria (i.e., Criteria 2 and 7). Criteria 2 and 7 are given equal weight because they are in, in fact, different dimensions of same element of risk, (i.e., ‘severity’).

The Codex Alimentarius Commission and other organizations that have proffered guidelines for food safety risk assessment have identified a number of fundamental criteria for the reliable conduct of such risk assessment. Common to most of these guidelines/recommendations is that the process should be “science-based”, and that the process used, including assumptions and limitations, should be “transparent”, i.e., that all data sources, assumptions, uncertainties and limitations and their consequences should be clearly and thoroughly documented as part of the risk assessments. These desiderata arise because perfect data to support risk assessment are seldom available, i.e., risk assessments usually represent only the best estimates available within the constraints and limitations of the data and current knowledge and those limitations, and their potential consequences for risk management decisions and the stakeholders affected by them, need to be made clear.

5. *Considering the various scenarios (described in Section 7.3 and Appendix N) to aggregate food-hazard pairs in order to identify the foods which should be identified as high-risk vs. not high-risk, which option(s) are more appropriate to consider and why?*

The risk rankings presented in Appendix N, based on various methods of aggregation, include results that are ‘unexpected’. The discussion included in Section N and the draft report also implicitly suggest that stakeholders, and perhaps even the model developers, consider that the rankings may not accurately reflect the true risks, e.g., on p. 2 of Appendix N it is stated “concerns were raised that perhaps there were potential biases in the model”.

While it is entirely possible for a risk assessment process to produce unexpected results, and for those unexpected results to be challenged by stakeholders and experts who hold different perceptions of relative risks, those model predictions may nonetheless be valid. The resolution of these apparent discrepancies should involve exploration of the logic of the model to determine whether the unexpected results can be understood (and represent scenarios and phenomena that may not yet have been perceived) or whether there are simply logical errors in the model.

Also, while the available epidemiological data are less than perfect, they do provide a point of reference by which to evaluate the credibility of the risk-ranking model - at least for product:hazard pairs that are part of the reporting/surveillance system. If the model rankings do not reflect the *known* situations for which reliable data are available, it suggests that there may be errors in the model.

An obvious problem with rankings based on the *sum* of the scores for each hazard considered to be associated with a food is that the attribution of more hazards, even if trivial, still adds to the score because some hazards have high scores for some criteria even if their likelihood of occurrence is negligible. Similarly, as noted earlier, the logic of the model doesn't have a way of effectively eliminating the risk from trivial hazards, e.g., due to cooking before eating, lethal treatments applied in hermetic packages, etc. As commented in the complementary review of the data upon which the RRM is based, using the total sum approach requires that every hazard needs to be assessed for every product. A specific example illustrating this is presented below).

Similarly, in Appendix N an averaging process was shown to distort the rankings because a food that has been associated with a hazard that is trivial, and another that has strong evidence as a persistent source of human illness will be under-estimated, even though it is logically clear that a real risk cannot be reduced by a lesser risk. The risk to human health from a single product is the sum of the risks from each associated hazard, even if some of the hazards are trivial.

The preferred option in Appendix N was to calculate the sum of the highest score for each criterion from any hazard considered to be associated with the product. ***This is a completely illogical approach and is not scientifically defensible*** because it involves mixing of attributes of one hazard with those of another, i.e., ***the approach completely undermines the logic that the model is based on*** and that such an approach should even be considered as a means of reconciling the model predictions with stakeholder expertise and perceptions suggests problems with the logic/structure of the model.

At least some of these problems are possibly due to incorrect approaches to combining variables that scale exponentially, and those that scale arithmetically, as more fully discussed in response to Q5a, below.

As an aside, at L. 20-22 on p. 107 of the draft report, it is stated that: “the risk ranking model might not capture the higher risk food-hazard pairs involving chemical hazards and undeclared allergens, because these agents do not grow in food and Criterion 4 is by default 0”. The basis of this comment requires some explanation, and link to the evidence base that suggests that the risk from these hazard groups is underestimated. However, examination of the risk ranking tables suggests that risk of allergens is frequently ranked very highly (see also my comments in the complementary data review), again suggesting that the basis of the comment needs further explanation/justification/investigation as to its validity. This issue is discussed further under Question 6b by reference to independent estimates of the relative burden of foodborne disease from different classes of hazards presented by WHO in late 2015.

a. Are there additional aggregation method(s) that might be considered? Please explain.

Yes.

As discussed earlier in this report, the risk ranking scheme involves combination of variables that scale exponentially, and the risk scoring process essentially produces a ranking on an exponential scale. However, if a product has risks attributable to multiple hazards, the overall risk is the ***sum*** of the risks, not the ***product*** of the individual risks. To explain, adding risk scores that are based on exponential scales is logically equivalent to multiplying the risks from each hazard, rather than adding them together to achieve an overall risk estimate. Given this, a more appropriate

approach to assessing the combined risk from multiple hazards is to convert the risk estimates to an arithmetic scale, add the risks together, and then convert the sum back to an exponential scale, i.e.,

Aggregate Risk Score, $RS_{total}F$ for food, F, with risk scores for each hazard given by RS_1F , RS_2F ... RS_nF for n identified hazards is:

$$RS_{total}F = \log_{10}(10^{RS_1F} + 10^{RS_2F} + 10^{RS_3F} + \dots + 10^{RS_nF})$$

As an example of why this approach is necessary, and one that will probably be familiar to people working in microbial food safety, consider a food that is initially contaminated with microorganisms, but is then further contaminated by contact with a contaminated surface. If the concentration of the organisms on the food is 10^2 CFU/g and the additional contamination is 10,000 (10^4) organisms, and the food weighs 100g, the final concentration is $(10^2 \text{ CFU/g} * 100\text{g} + 10^4 \text{ CFU})/100\text{g} = 20,000/100\text{g} = \log_{10}2.3\text{CFU/g}$. i.e., despite that the additional cross contamination is the equivalent of 10^2 CFU/g, the final contamination is *not* 10^{2+2} CFU/g but 10^2+10^2 CFU/g.

Using this (more) logically defensible approach, the cumulative risk score would scale more naturally, and will be less affected by differences in the total number of hazards ascribed to the various foods.

An overt logical error was discovered in the model, however, relating to aggregation irrespective of the treatment or weighting applied that further compounds the anomalies.

To illustrate the problem, in the assessment of the microbial risks associated with finfish, 16 microbial hazards were identified and evaluated in the risk ranking, leading to a relative risk of ~400. In a separate category 'finfish associated with histamine' was associated with only one hazard, i.e., histamine, and generated a risk score of only 25. This is illogical. *Every other microbial hazard associated with finfish can also be associated with finfish capable of developing high histamine levels.* Epidemiological evidence shows that histamine intoxication from scombroid fish, while usually not severe, is the commonest cause of illness related to finfish. As such, the risk from histamine producing finfish is systematically under-estimated because of this (arbitrary?) association of hazards with different types of finfish, as was suggested above.

More importantly, however, this treatment of the hazards data reveals a systematic logical error in the model that was alluded to (in the discussion above) about the cumulative effect of multiple hazards. In the scoring, every individual hazard score includes a score for 'consumption'. Therefore, every additional hazard deemed to be associated with a product adds to the 'consumption' score again, even though the presence of an additional hazard does not change the consumption, it simply means more hazards are associated with the *same amount* of consumption. In the logic of the scoring scheme, however, each additional hazard effectively assumes that the consumption is increased. In the case of finfish, with a consumption score of '3', the sixteen hazards identified as being associated with finfish mean that the contribution of consumption is scored as 48 instead of 3, on a scale where the maximum consumption score should be '9'. Presumably, this error is inherent throughout the relative risk scores where more than one hazard is associated with a food or food category. As such, the more hazards deemed to

be associated with the hazard, the more inflated the consumption score becomes and the more distorted the relative ranking becomes, confirming comments above about a correct method of aggregation of scores from multiple hazards. Currently, the method is logically incorrect. Similarly, where the 'at risk' population is only a subset of the population, the 'consumption' values have to be modified accordingly (as was discussed in the complementary report on the model data).

6. Given the underlying data supporting the scoring, what are the considerations to take into account when identifying high risk vs. not high risk food-hazard pairs or foods?

There are numerous factors that influence food-borne risk, as discussed above in response to Question 1. These include:

- the potential for contamination with a hazard at a level likely to cause human illness, or the potential during normal handling to increase (grow) to a level that could cause illness (considering also the proportion of the population susceptible to the hazard),
- the normal use of the product (e.g., RTE or cooked before consumption),
- the existence of reliable CCPs for those hazards and their reliable implementation, or quality assurance systems that reliably detect contamination and thereby effect removal of contaminated lots from distribution,
- whether the product can be recontaminated with a disease-causing dose *after* a reliable CCP treatment/action or quality assurance process,
- the severity of the symptoms associated with disease caused by consumption of the hazard,
- the frequency of consumption of the food (whether by individuals, or total consumption by population, or only by the susceptible population, as relevant),
- whether the hazard accumulates in the body of the consumer or whether each exposure is a discrete event (and whether earlier exposures provide protection against subsequent exposures, i.e., immunity, or predispose the consumer to more severe symptoms upon subsequent exposure, e.g., induced hypersensitivity), etc.

As noted earlier, all these elements of risk are implicit (though perhaps confounded) in the six factors nominated for inclusion by FDA, and are translated in various combinations into the seven criteria adopted for the draft Risk Ranking Model (RRM). Consumption was not explicitly considered in the six factors identified by FDA, however, but that omission may be appropriate if the risk is intended to be ranked on a per-serving basis. The likelihood of a disease-causing dose also is not considered explicitly (nor implicitly as far as I can determine) and this seems to be a weakness in the draft RRM.

Interestingly, the FDA's decision criteria/logic for identification of Potentially Hazardous Foods/Temperature Controlled for Safety Foods seems to be relevant to identification of "riskier" foods (i.e., at least those subject to potential microbial contamination). Thus, that evaluation system might have provided a useful and consistent starting point for the RRM which is intended to identify "high-risk foods that require additional record keeping".

a. *What are the pros and cons in establishing a threshold considering all three types of hazards in the model, vs. drawing a line separately for microbial, chemical and undeclared allergens in foods?*

Ideally, there would be a simple, transparent and objective risk evaluation system that accurately estimates, and ranks, the overall risk from a diversity of microbial, chemical and allergen hazards that could contaminate the identified foods of interest. However, the variety of food composition, processing and handling steps, the responses of the hazards to those treatments, including over time, the type of effects of the hazards on consumers, and the strong differences in consumer susceptibility, seem to preclude such a system being feasible, unless supported by much more data than is currently available to assess each of the relevant factors.

As suggested in earlier comments there appear to be anomalies in the preliminary risk rankings presented, with estimated risk from undeclared allergens appearing to be dominant despite the lack of confirming, or even supporting, epidemiological evidence. However, 'absence of evidence is not evidence of absence' and there may be a high burden of (unreported) illness from allergens in the U.S. food supply. The same is true of chemical hazards in the U.S. food supply, i.e., there is a paucity of relevant data.

However, while it was not available at the time of preparation of the draft RRM, the World Health Organization report on the Global Burden of Foodborne Disease (released in December 2015; http://www.who.int/foodsafety/areas_work/foodborne-diseases/ferg/en/) provides an independent, expert and well-resourced evaluation of relative risk from different types of hazards in foods. The WHO report considers the burden of foodborne illness in different regions of the world including the 'AMR' region comprising Canada, U.S.A. and Cuba. The population of Cuba is ~11 million. The population of Canada is ~36 million. The population of U.S.A. is 323 million. As such, the estimated burden of disease in AMR will be dominated by U.S. statistics.

Collectively, Tables A8.4, A8.6, and A8.7 in the WHO report permit estimates of the disease burden from selected chemicals, peanut allergens, and a plethora of microbial (bacterial, protozoal and fungal pathogens and their toxins) hazards. Due to data limitations, however, the only 'chemical' hazards considered were aflatoxin, dioxin and cyanide from Cassava. Clearly, there is negligible risk from Casava consumption in North America.

Table A.6 in the WHO report considers the burden of foodborne illness in: i) the north America Region, ii) European region and iii) Western Pacific Region (a mix of 37 countries including affluent 'westernized' nations and developing island nations). In those regions the combined burden from peanut allergens was ~10% of the burden from dioxin and aflatoxin combined. Given that the Western Pacific region includes many tropical nations, the relative risk from aflatoxin in those countries may be relatively higher, thereby reducing the apparent relative importance of peanut allergens.

Nonetheless, the estimated DALY burden per 100,000 population in those three regions for aflatoxin and dioxin combined was 2 (95% CI: 0.6-24), while the DALY burden for all microbial hazards was 51 (95% CI: 41-112). Based on the statistic above, the relative DALY burden per 100,000 for peanut allergens would be ~0.2. (Peanut allergies are considered to be the most common in the 'developed' world). These relative burden of disease estimates do not seem to accord with the relative risks identified in the draft model for review, suggesting that relative

risks attributed to microbial hazards, chemical hazards and allergens in the draft model are not consistent. The discrepancy between risk estimates for allergens in the draft RRM and Minor *et al.* (2015, Risk Anal., 35:1135-1139) was noted earlier. It is noted that chemical hazards and allergens have separate scoring schemes, especially for severity, and that relative susceptibility (or proportion of the population at risk) is not explicitly considered in the model.

Given these considerations and the apparent discord between the WHO estimates and those from the draft RRM, it is this reviewer's opinion that separate thresholds are needed if using the current RRM for determination of foods that require additional record-keeping due to possible high risk of: i) microbial hazards, ii) chemical hazards and iii) allergens. Further data are needed to reliably determine relative risks from these three categories of hazard. Alternatively, revision of the criteria scores for each type of hazard may be required to generate risk estimates that are comparable on a burden-of-disease basis.

7. *Are the seven criteria and scoring definitions implemented appropriately in the SAS codes (Appendix P) and the Access Model?*

I do not have sufficient expertise in SAS and R, nor do I have sufficient expertise in M/S Access, to be able to make informed comment about the code underlying the RRM interface.

- a. *Does the scoring logic described in Section 4 Figures 4-1 to 4-7) appropriately represent the scoring definitions described for each of the criteria in Section 3? If not, please describe what changes need to be made to correct it.***

Note. The criteria definitions were described in Section 2. Section 3 was concerned with identification of product:hazard pairs. Section 4 described the criteria definitions again, together with their implementation.

Criterion 1, Section 2.2.1 text agrees with the scheme described in Figure 4-1.

Criterion 2, Section 2.2.2 text agrees with the scheme described in Figure 4-2.

Criterion 3, Section 2.2.3 text agrees with the scheme described in Figure 4-3.

Criterion 4, Section 2.2.4. The text does not agree with the data/text in Figure 4-4 because no scores are given for the growth potential descriptors and shelf life descriptors at Section 2.2.4, while they are defined in Figure 4-4. Perhaps more importantly, the logic for combination of the two scores into a single score of 1, 3 or 9 is not described, i.e., there is a lack of transparency. Importantly, if the two scores are added, it appears that a value >10 is scored as 9, a value between 6 and 10 is scored as 3, and if the value is between 2 and 4 the value is scored as 1. If the values are multiplied, then a score (i.e., product) >27 and up to 81 is scored as '9'. Values of 9 are scored as 3 and values of 1 to 3 are scored as 1. By either approach, the relativities of the scale (and the risk contributions) are distorted by this process. The rationale for combination of the scores should be described clearly in the document.

Criterion 5, Section 2.2.5. Same comments as above for Criterion 4 and Figure 4-4, i.e., no scores are given for the 'contamination probability during manufacturing' descriptors and 'steps taken to reduce contamination' descriptors at Section 2.2.4 but they are given in Figure 4-4.

Perhaps more importantly, the logic for combination of the two scores into a single score of 1, 3 or 9 is not described, i.e., lack of transparency. The rationale for combination of the scores should be described clearly in the document.

Criterion 6, Section 2.2.6 text does not agree with Figure 4-6 because Figure 4-6 includes scores for expert elicitation results that are not discussed at all in Section 2.2.6. Section 4.2.6 does mention that expert elicitation was used where specific data were not available from NHANES. Thus, the text at Section 2.2.6 is not fully consistent with Section 4.2.6.

Criterion 7, Section 2.2.7 text does not agree with Figure 4-7 because Figure 4-7 includes scores from 'expert opinion' that are not discussed at all in Section 2.2.7. Section 4.2.7 does mention that expert elicitation was used where specific data were not available from other sources (i.e., Minor, 2015; Scharf, 2011). Thus, the text at Section 2.2.7 is not fully consistent with Section 4.2.7.

b. Are the scoring logic and order of preference accurately implemented in the SAS codes? (Please select 2-3 out of the 7 criteria for this evaluation). If not, please specify what changes need to be made.

I do not have sufficient expertise in M/S Access to be able to make informed comment about the accuracy of the SAS code underlying the RRM interface. However, I used the model in Access to generate ranking data for selected foods and hazard sets. By using the "Adjust Scores" dialog box I was able to identify the scores for each criterion for each hazard associated with the selected food and, from this, I was able to check the simple additions (Eqn. 3). I then changed weighting combinations and re-ran the scenario and rechecked the additions leading to the FRS. I repeated this process for another set of weighting combinations. From this I was able to check the correctness of implementation of the weighting factors.

The food:hazard pairs I considered were:

- i) "Eggs" for all hazards and, from the four outputs, I selected "Egg dishes" (which encompassed 7 hazards) for further assessment by checking additions leading to the FRS, and additions after changing the weighting.
- ii) "Pasta -Dried Pasta" and "All Hazards", which encompassed 8 associated hazards.
- iii) "Seafood - Finfish" for "Microbial Hazards" only and from the output selected "Finfish" but also "Finfish (Histamine Producing Species)" for more detailed assessment. For "Finfish" there were 16 discrete hazards, while for 'Finfish (Histamine Producing Species) there was only one hazard, i.e., histamine.

In all cases, the additions and aggregated scores were correct and consistent with correct implementation of Equation 3. However, this assessment did reveal a logical error in the model that was described in response to Question 5a (regarding accumulation of consumption scores).

Also, from data presented in the "Adjust Scores" dialog box the calculation of the uncertainty score was 'checked'. The uncertainty score should be the simple sum of the individual uncertainty scores for each criterion for each hazard associated with the food. The definition of the uncertainty score in Section 4.4.2 and Equation 4 does not describe inclusion of weighting factors, but it was noted (through the 'experiments' described above) that the Uncertainty Score (and the

Confidence Score) are also affected by the designation of the weighting factors specified in the 'Ranking Criteria and Weights' in the same way (algebraically) that the criterion scores are affected. ***Whether such weighting of the FCS and FUS is logically justified is debatable, but it certainly needs to be described in the documentation.*** Whether the omission of weighting factors in the definition of Eqn.4 is an oversight, or whether there is an error in the model code is not clear but a correction is required in one or the other.

As an aside, there seems little value in including both the FCS and FUS as model outputs, as they effectively measure the same properties of the risk-ranking scores but on inverse scales, i.e., if the FSC for a criterion is 9, then the FUS must be 1, and *vice versa*, and when FCS is 3 so is the FUS. The net effect of this is that the two scores are strongly negatively correlated. To illustrate this, an approximation between these 'scores' is given by:

$$\text{FCS} = 64.85 * \text{number of hazards associated with food} - \text{FUS} \text{ (Eqn. Rev1)}$$

The observed FCS is plotted against the predicted FCS based on Eqn. Rev1 in Figure B, below.

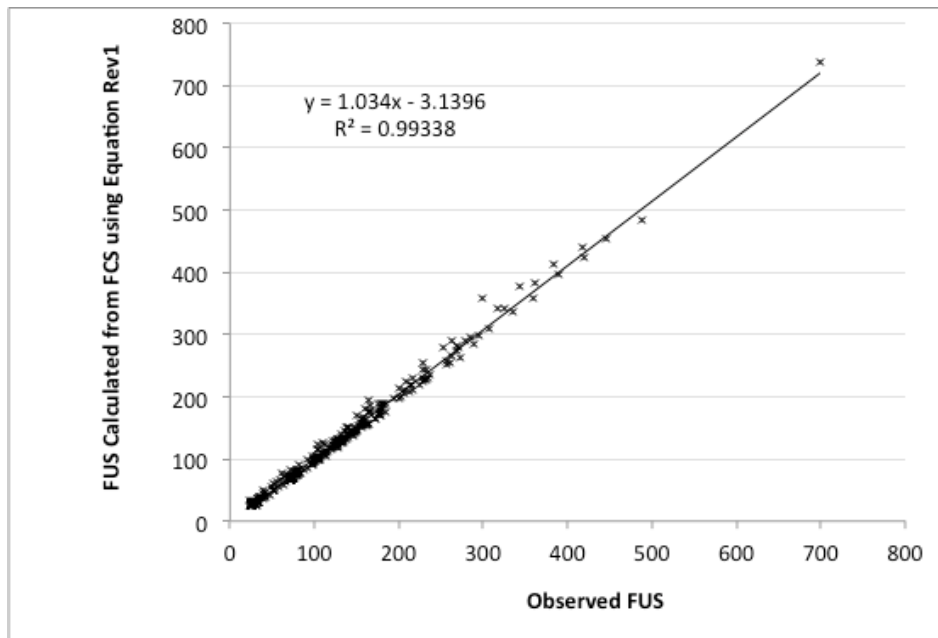


Figure B. Results of a simple relationship (Eqn. Rev1) relating FCS to FUS, and showing the success of that relationship (regression and r^2 value shown) in predicting FUS from the FCS.

On the basis of this analysis, the additional insights offered by the inclusion of both FSC and FUS in the model outputs should be explained.

c. Are equations 1 through 4 and data weighting factors accurately implemented in the model (either the SAS codes or the Access Model)?

I do not have sufficient expertise in M/S Access to be able to make informed comment about the accuracy of the code underlying the RRM interface, nor whether the code accurately reproduces the logic of the model as described in Section 4 of the draft report. However, see response to Question 7b that discusses assessment of the correctness of implementation of Equation 3 and Equation 4.

8. In the Access Model, is the underlying relational database including lookup tables and algorithm appropriately designed and implemented? If not, please explain what changes should be considered.

I do not have sufficient expertise in M/S Access to be able to make informed comment about the accuracy of the code underlying the RRM interface as it applies to analysis of the data in the look-up tables and whether that has been implemented as described in the draft report.

9. Is the user interface of the Access Model sufficiently described for the user to understand each component of the model, e.g., foods, hazards, ranking criteria, results, and cited references?

The User Guide still refers to the model as the HRF, despite noting (in the Introduction) that the name for the model is now the Risk Ranking Model for Product Tracing (RRM-PT). This is a trivial point, but it is unnecessarily confusing, particularly given that document could have been made 'up to date' with a few minutes of 'search and replace' in Microsoft Word.

I tried to read the User-Guide and use the RRM-PT model interface as though I *hadn't* spent many days going through the draft report in detail, i.e., to consider both from the perspective of an uninitiated potential user. From that I consider that the user-interface is not sufficiently well-described for a user to be able to use the model appropriately and to be able to correctly interpret the outputs.

The 'User Guide (short version)' is necessary, but even that is insufficient because it assumes that the reader/user has an understanding of the structure and logic of the risk ranking model. The User Guide (Introduction) refers to the quite detailed Sections 2 and 4 of the Draft Report. The user "help" functions in the interface itself are useful, but not of themselves adequate to enable a user to correctly use the model and to interpret its outputs. It would improve the User's Guide if a few paragraphs were added to the Introduction that explained the basis of the seven criteria, the scoring scales used (including some mention of underlying data sources), how they are combined to generate a score for each food:hazard pair, and how those scores are aggregated to generate an overall risk score. The use of terms like 'scenario', 'criteria', 'repository', 'weights' etc. needs to be explained in terms of their use in the RRM-PT.

In summary, as someone who (now) understands the basis of the RRM and its supporting data and algorithms, I found the user manual very useful to enable me to begin to use the model to generate risk ranking estimates. Without the User's Guide Short Version I don't think I would have been able to do that intuitively, even having thoroughly read and worked through the draft report, suggesting that the user interface is not 'stand-alone' for novice users.

Other comments on the interface and user guide:

The RRM-PT interface suggests that each criterion value can be traced to a data source or expert opinion source. I followed many of the 'literature' links and the resulting screen 'dialog-box' contained no information. This requires some explanation in the user guide and the interface itself. Also, sometimes a 'literature' hyper-link led only to description of the data as 'expert opinion'. This aspect seems inconsistent and conveys the sense that RRM is still incomplete.

The User Guide refers to Table 1, but Table 1 is apparently not included in the Users Guide. P3 (User Guide), third dot point about "View Baseline Data". I think the intended word is 'underpinning' not 'underlining'?

In the dialog box about (re)assigning scores to composite score criteria, the help function doesn't explain that a password is required, nor how to obtain it.

P8 (User Guide) dot point 4, second line.. There is an error in the text "we be used...", but I'm not sure what the intended text is.

10. How often should the model be updated, considering the data sources and data currently available and types of data that might become available in the future?

The following is an extension of comments from my complementary review and report on the data behind the RR model:

I am not sufficiently expert to be able to offer a reliable response to this question but would suggest that the rate of change of food processes and product formulation might mean that the model and Criterion values would need to be reviewed every 3 to 5 years. However, if it becomes evident that radically different processes or products are introduced, or products are sourced from new/different suppliers, it would be prudent to evaluate *before introduction of those products* whether those changes introduce a different level of public health risk. Nonetheless, the evaluation of the model presented here suggests that the current draft RRM is not yet ready for practical use and risk management decision-making.

11. Is the draft report clear in its description of the risk ranking approach, criteria and scoring definitions, and model limitations? If not, please identify which aspects are unclear or could be more transparent.

Comments about the transparency of the process and the model, identification and documentation of relevant and authoritative sources, and limitations in the data and the model were made above and also in the complementary report and review of the data. Specific comments are presented in Section III, below. Further, there appear to be many minor presentation errors and examples of use of jargon and idioms that may not have unambiguous meaning to all readers.

12. Do you have any additional comments? Please share them in your review.

No. All relevant comments have been made above.

III. SPECIFIC OBSERVATIONS ON DRAFT REPORT FOR PEER REVIEW: RISK RANKING MODEL FOR PRODUCT TRACING AS REQUIRED BY SECTION 204 OF FSMA (RRM-PT Draft Report) WITHIN THE CONTEXT OF THE MODEL ITSELF.

Page	Paragraph/Line	Comment
General		The phrase “in order to” can in almost all circumstances be reduced to “to” without any loss of inference or meaning. “In order to” is frequently used in the document and can be simplified to “to”
P13	L1	‘scoring’ should be ‘score’
P15	L9	‘in’ should be ‘for’
P15	L18	‘issue’ should be ‘issues’
P16	L2	‘outbreak’ should be ‘outbreaks’
P16	L3	‘representing’ should be ‘represent’
P17	L16	‘definition’ should be ‘definitions’
P19	L5	delete ‘or’
P19	L37	delete ‘foods’ after ‘ready-to-eat’
P19-P20	L40-L2	This sentence essentially repeats the previous sentence (i.e., P19, L36 – 39).
P20	L12	font size is inconsistent
P21	L24	primary production infers, ‘on the farm’, i.e. ‘on the farm’ is a tautology.
P24	L8	delete ‘be’
P25	L5	‘pair’ should be ‘pairs’
P25	L18	‘examples’ should be ‘example’
P26	L2	delete second ‘in’
P26	L13	change ‘identified’ to ‘identify’
P28	L6	‘polynuclear’ should be ‘polycyclic’
P30	last para in right-hand column	delete ‘and’ in first line of para
P31	L6	correct to ‘to assign a numerical score of...’
P31	L10	correct ‘qualitatively’ to ‘qualitative’
P33	L5	change ‘detail’ to ‘detailed’
P37	L5	insert ‘that’ after ‘address’; change ‘pair’ to ‘pairs’
P40	L12	change ‘hazard’ to ‘hazards’
P44	Figure 4-3	In the second table in the right-hand column, the instructions should say “For each prevalence study, assign the data weight”, i.e., not assign the geographic weight.
P47	L26	delete ‘to’
P48	L1	‘there’ should be ‘those’
P48	L17	the phrase ‘did not have non-zero quantitative prevalence...’ is convoluted and I’m still not exactly sure what it means. Try to reword more clearly.
P48	L30	‘detection’ should be ‘detections’
P51	L6	insert ‘is’ after ‘contamination’
P52	L17	insert ‘and’ before ‘chemical hazards...’
P56	L27	change ‘step’ to ‘steps’
P64	L6	change ‘is’ to ‘as’, or delete ‘is’
P72	L25	FDA has expert opinions? (Perhaps change to ‘opinions’?)
P77	L6	insert ‘the’ after ‘run’
P77	L29	the data don’t underline the model, they ‘underpin’ it

Page	Paragraph/Line	Comment
P84	L12	delete “number of”
P91	L19	change ‘give’ to ‘gives’
P96	L10	change ‘frequently’ to ‘frequent’
P96	L13	change ‘precipitate’ to ‘precipitous’; However, either term is a subjective expression that seems largely to describe a natural consequence of the scoring scheme, more than any real phenomenon related to food-hazard pairs. See also relevant discussion about this “observation” in response to Question 4.
P103	L13	insert ‘in’ before ‘Table 7-4’
P103	L19	insert ‘; with’ before ‘highly ranked..’
P103	L29 (twice), L30, L31, L39	change ‘has’ to ‘have’
P104	L2, L5	change ‘has’ to ‘have’
P107	L14	delete ‘a’ before ‘9’
P107	L22	insert ‘on’ before ‘option 4.’
P107	L26	insert ‘on’ at end of line after ‘analysis’
P110-111	L1-L37	the term ‘a greater degree of’ can more easily be expressed as ‘more’, i.e., there were more data gaps....
P111	L20	change ‘chose’ to ‘choose’
P122	L4 ff.	The bibliographic details are incomplete. Minor et al (2015) is now published, in Risk Analysis, 35, pp. 1135-1139.

IV. SPECIFIC OBSERVATIONS ON APPENDICES TO THE DRAFT REPORT FOR PEER REVIEW: RISK RANKING MODEL FOR PRODUCT TRACING AS REQUIRED BY SECTION 204 OF FSMA (RRM-PT Draft Report) WITHIN THE CONTEXT OF THE MODEL ITSELF.

Appendix	Page/Row	Paragraph/Line/Column	Comment
			See responses to specific questions

V. SPECIFIC OBSERVATIONS ON RISK RANKING MODEL FOR PRODUCT TRACING: USER’S GUIDE SHORT VERSION.

Page	Paragraph/Line	Comment
		See responses to specific questions, in particular Q. 9.

VI. SPECIFIC OBSERVATIONS – PROVIDE SPECIFIC OBSERVATIONS, CORRECTIONS, OR COMMENTS ON THE ACCESS DATABASE FDA’S HIGH RISK FOODS (HRF) MODEL.

Menu Choice	Tab	Steps taken within the tab	Comment
			See responses to specific Questions, in particular Q. 9.

Reviewer #5

Peer Review Comments on FDA's Draft Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA:

HRF Model, the HRF Model Report, the HRF Model User's Guide, and the HRF Model Design Report

Reviewer #5 – **To be completed**

I. GENERAL IMPRESSIONS

II. RESPONSE TO CHARGE QUESTIONS

1. *The draft risk ranking model uses the FSMA statutory factors to define seven criteria for scoring food-hazard pairs.*
 - a. *Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.*
 - b. *Within the bounds of the FSMA-mandated factors, are there additional criteria beyond the seven criteria that should be considered? If so, please describe these additional criteria that might be considered and why.*

2. *Are the scoring definitions for all criteria appropriate?*
 - a. *Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and undeclared allergens)? If not, please describe changes that might be considered and why.*

 - b. *Is the value function 0-1-3-9 and scoring matrix appropriate for the intended purpose to inform the designation of high-risk foods? If not, please describe changes that might be considered and why.*

3. *Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.*

4. *Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider*

for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.

5. *Considering the various scenarios (described in Section 7.3 and Appendix N) to aggregate food-hazard pairs in order to identify the foods which should be identified as high-risk vs. not high-risk, which option(s) are more appropriate to consider and why? Are there additional aggregation method(s) that might be considered? Please explain.*

6. *Given the underlying data supporting the scoring, what are the considerations to take into account when identifying high risk vs. not high risk food-hazard pairs or foods? What are the pros and cons in establishing a threshold considering all three types of hazards in the model, vs. drawing a line separately for microbial, chemical and undeclared allergens in foods?*

7. *Are the seven criteria and scoring definitions implemented appropriately in the SAS codes (Appendix P) and the Access Model?*
 - a. *Does the scoring logic described in Section 4 Figures 4-1 to 4-7) appropriately represent the scoring definitions described for each of the criteria in Section 3? If not, please describe what changes need to be made to correct it.*
 - b. *Are the scoring logic and order of preference accurately implemented in the SAS codes? (Please select 2-3 out of the 7 criteria for this evaluation). If not, please specify what changes need to be made.*
 - c. *Are equations 1 through 4 and data weighting factors accurately implemented in the model (either the SAS codes or the Access Model)?*

8. *In the Access Model, is the underlying relational database including lookup tables and algorithm appropriately designed and implemented? If not, please explain what changes should be considered.*

9. *Is the user interface of the Access Model sufficiently described for the user to understand each component of the model, e.g., foods, hazards, ranking criteria, results, and cited references?*

10. *How often should the model be updated, considering the data sources and data currently available and types of data that might become available in the future?*

11. *Is the draft report clear in its description of the risk ranking approach, criteria and scoring definitions, and model limitations? If not, please identify which aspects are unclear or could be more transparent.*

12. *Do you have any additional comments? Please share them in your review.*

III. SPECIFIC OBSERVATIONS ON DRAFT REPORT FOR PEER REVIEW: RISK RANKING MODEL FOR PRODUCT TRACING AS REQUIRED BY SECTION 204 OF FSMA (RRM-PT Draft Report) WITHIN THE CONTEXT OF THE MODEL ITSELF.

Page	Paragraph/Line	Comment

IV. SPECIFIC OBSERVATIONS ON APPENDICES TO THE DRAFT REPORT FOR PEER REVIEW: RISK RANKING MODEL FOR PRODUCT TRACING AS REQUIRED BY SECTION 204 OF FSMA (RRM-PT Draft Report) WITHIN THE CONTEXT OF THE MODEL ITSELF.

Appendix	Page/Row	Paragraph/Line/Column	Comment

V. SPECIFIC OBSERVATIONS ON RISK RANKING MODEL FOR PRODUCT TRACING: USER'S GUIDE SHORT VERSION.

Page	Paragraph/Line	Comment

Page	Paragraph/Line	Comment

VI. SPECIFIC OBSERVATIONS – PROVIDE SPECIFIC OBSERVATIONS, CORRECTIONS, OR COMMENTS ON THE ACCESS DATABASE *FDA'S HIGH RISK FOODS (HRF) MODEL*.

Menu Choice	Tab	Steps taken within the tab	Comment

IV. PEER REVIEWER COMMENT TABLE

I. General Impressions		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	<p>Despite the complex nature of connecting multiple factors and a myriad of data sources to construct the high risk foods model, the current draft model is straightforward, easily understood and generally logical. However, the simplicity of the model with seven criteria that are proxy for potential exposure and risk of illnesses from “high risk foods”, necessitates rigorous scrutiny of the underlying data and process that were relied upon to inform the scoring of each of the seven criteria for each food-hazard pair, as well as how the seven criteria are combined and how food-hazard pairs are integrated to index “high risk foods”. It is this reviewer’s understanding that examination of the underlying data is being undertaken by another peer review panel, hence the comments provided herein focus on the seven criteria scoring method and integration of the seven criteria scores within a food-hazard pair as well as integration across multiple hazards for a given food. Overall, it is this reviewer’s impression that scoring and underlying data supporting scoring for each of the seven criteria have been thoroughly examined by the Agency in the development of the current model. However, the process of integrating scores for each food-hazard pair and integrating across hazards for each food is still work in progress, and further sensitive assessment is needed to assure that the method selected for the final model is sound and supported by the most robust science/policy rationale. It is also noted that significant progress has been made to adequately capture chemical risks in this current draft model. Some degree of unbalanced emphasis on microbial risks remain and are noted below in response to the various charge questions of this peer review.</p>	
Reviewer #2	<p>I highly welcome the initiative of FDA to develop the High Risk Foods Model and Associated Food-Hazard Combinations-Model risk ranking model and support the efforts towards a more systematic approach for risk ranking. I wish to congratulate the project team for the huge work done to collect the needed data for 1286 food-hazard combinations. I believe that the initiative that includes both chemical and microbial hazards will be successful and recognized by the other food safety agencies that need a coherent and pragmatic way to rank risk associated with food.</p> <p>However, as a general remark, I would like to stress the need of better inclusion of uncertainty in the risk ranking process. Because of the high numbers of ranked objects, it is illusory to expect that all the criteria used will be accurately and precisely evaluated for all the food-hazard combinations.</p> <p>MCDA techniques and methods are useful to overcome these difficulties. The use of ordinal scale to score the different criteria does not allow the use of simple aggregation methods such as weighted sum. It is relatively easy to show the possible errors of the used weighted sum. Indeed, as FDA developed a sound tool for risk ranking (i-Risk) it will be easy to run this tool for a limited number of food-hazard combinations for which quantitative data are available, and then compare the ranking order obtained with i-Risk and the one obtained with the reported model.</p>	

I. General Impressions		
REVIEWER	COMMENT	RESPONSE
	<p>To overcome the problem linked to the uncertainty about the scoring of the different criteria and the ordinal scale problem other MCDA techniques need to be deployed. One promising method is ELECTRE III.</p>	
Reviewer #3	<p>Mitigating outbreaks of foodborne illness requires a consideration of the various risks associated with many different foods and food byproducts. The Food and Drug Administration (FDA) has developed a semi-quantitative risk ranking model that integrates several criteria and metrics associated with food assessment. The document offers a robust approach for the purpose of risk ranking that utilizes a multi-criteria scoring model to derive a risk score and ultimately inform decisions on food-based health risks.</p> <p>Within this framework, the general idea is that foods with more elevated risk scores should be a higher priority for consideration. The use of multi-criteria models in risk prioritization are particularly suitable when it is impractical to build and populate a full causal risk model, so where multi-criteria models instead act as proxies that are measurable and which are thought to be associated with risk. The proposed approach integrates significant volume of historic and measurable data with expert judgment.</p> <p>With this said, I have some concerns about specific algorithms used. It seems the risk factors should be multiplicative because they form a chain of events leading to consequences, but the document utilizes additive model. Additive model may be a reasonable approximation, but additional clarification and justification may be required. In any event, there are no easy shortcuts to producing a score that has the precision of a structured risk analysis and the proposed approach is a major improvement over traditionally used dashboards or individual risk indicators.</p>	
Reviewer #4	<p>Detailed exploration of the draft model and its accompanying documentation suggests that there are anomalies in the estimates produced by the model compared to what would be expected from the available epidemiological data.</p> <p>At least some of these anomalies seem to arise from logical errors in the model or that the model does not accurately represent the mathematical relationships between risk-affecting factors. For example, some of the problems seem to arise from inappropriate combinations of factors that scale on algebraic scales and those that naturally scale exponentially. In the consideration of methods for aggregation, one proposed approach cannot be justified logically. Also, all aggregation approaches introduce a logical error based on erroneous amplification of consumption scores, and there seems to be an anomaly in the implementation of the weighting factors as they apply to uncertainty and confidence scores.</p> <p>In general, the apparent problems in the model seem to arise from trying to use a semi-quantitative risk matrix approach. The simplification of risk-affecting factors that are described by continuous variables</p>	

I. General Impressions		
REVIEWER	COMMENT	RESPONSE
	<p>into sets of three (or four) categorical values, introduces demonstrable anomalies in the scoring and distorts the relative risk estimates. Given that the RRM is developed in computer software that is capable of complex computations, and that there is no apparent need to simplify the calculations, e.g., if the model needed to be able to be used by people without access to a computer, the reasons for this approach are unclear. A better outcome would be expected by preserving the maximum value in the (quantitative) data and building a risk estimation model that achieves that.</p> <p>In places the discussion in the report seems more about the mechanics of the modelling process and interface function, rather than in the underlying logic that leads to model predictions and evaluation of their credibility.</p> <p>The report itself contains many minor presentation errors.</p> <p>In short, the errors apparent in the model and in the accompanying documentation suggest that the model is not yet ready for its intended use.</p>	
Reviewer #5	[To be completed]	

II. Response to Charge Questions

CHARGE QUESTION 1: *The draft risk ranking model uses the FSMA statutory factors to define seven criteria for scoring food-hazard pairs.*

CHARGE QUESTION 1.a: *Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.*

REVIEWER	COMMENT	RESPONSE
Reviewer #1	<p>The seven criteria in the draft model are appropriate for a multi-criteria decision approach. These criteria capture the essence of the FSMA factors; each FSMA factor was represented by two criteria, except for FSMA (iii) and (iv) which were represented by a single criterion 5. The seven criteria are proxy for exposure and risks, hence, appropriate for the purpose of a risk-based ranking model tool.</p> <p>The emphasis of FSMA on manufacturing aspects (two FSMA factors), but represented by just one criterion 5 should be noted and considered in the weighting of the seven criteria in the aggregation across food-hazard pairs to derive a composite score of a single food (more later).</p>	

CHARGE QUESTION 1.a: Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.		
REVIEWER	COMMENT	RESPONSE
	<p>(i) the known safety risks of a particular food, including the history and severity of foodborne illness outbreaks attributed to such food, taking into consideration foodborne illness data collected by the Centers for Disease Control and Prevention;</p> <ul style="list-style-type: none"> • criteria 1 (frequency and occurrence of outbreaks) • criteria 2 (severity of illness) <p>(ii) the likelihood that a particular food has a high potential risk for microbiological or chemical contamination or would support the growth of pathogenic microorganisms due to the nature of the food or the processes used to produce such food;</p> <ul style="list-style-type: none"> • criteria 3 (likelihood of contamination) • criteria 4 (growth potential) <p>(iii) the point in the manufacturing process of the food where contamination is most likely to occur;</p> <ul style="list-style-type: none"> • criteria 5 (manufacturing process/contamination intervention) <p>(iv) the likelihood of contamination and steps taken during the manufacturing process to reduce the possibility of contamination</p> <ul style="list-style-type: none"> • criteria 5 (manufacturing process/contamination intervention) <p>(v) the likelihood that consuming a particular food will result in a foodborne illness due to contamination of the food;</p> <ul style="list-style-type: none"> • criteria 6 (consumption) • criteria 3 (likelihood of contamination) <p>(vi) the likely or known severity, including health and economic impacts, of a foodborne illness attributed to a particular food.</p> <ul style="list-style-type: none"> • criteria 7 (economic impact) • criteria 2 (severity of illness) 	
Reviewer #2	<p>Multi-criteria decision analysis/approach (MCDA) in general follows the sequence below:</p> <ul style="list-style-type: none"> - Identifying objectives - Identifying options/alternatives for achieving the objectives - Identifying the criteria to be used to compare the options 	

CHARGE QUESTION 1.a: Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.		
REVIEWER	COMMENT	RESPONSE
	<ul style="list-style-type: none"> - Analysis of the options - Making choices, and - Feedback <p>To adapt the MCDA to risk ranking we can define the sequence as following:</p> <ul style="list-style-type: none"> - Identify the objective: health impact that can be directly linked with food consumption (overall risk) - Identify the list of food-hazard combination that contribute to the overall risk - Identifying the criteria to be used to compare the contribution of the different food-hazard combinations to the overall risk - Collect data for each food-hazard combinations relative to the identified criteria - Aggregate the different criteria and rank the food-hazard combinations - Continuous reassessment of the choices made in the past... <p>To answer to the question “<i>Are the seven criteria used in the draft model appropriate for a multicriteria decision approach?</i>” the objectives need to be clearly defined.</p> <p>Multi-criteria analysis, in the current project, consists on the identification of food-hazard combinations that contribute most to the foodborne burden: risk ranking. The risk ranking has to be based on measurable criteria to assess the extent to which each combination contributes to the overall burden or risk.</p> <p>The chosen criteria are: Criterion 1: Frequency of outbreaks and occurrence of illnesses (i.e., Epidemiological Link) Criterion 2: Severity of illness, taking into account illness duration, hospitalization and mortality Criterion 3: Likelihood of contamination Criterion 4: Growth potential/shelf life Criterion 5: Manufacturing process contamination probability/intervention Criterion 6: Consumption Criterion 7: Economic impact</p> <p>Criterion 1: This criterion is appropriate and relevant to distinguish between a high risk food-hazard high combination and low risk one in this ranking problem. However, this criterion is applicable only if data for all the food-hazard combinations are available and it has not the same interpretation for all the hazards: total number of cases including or not sporadic cases. Considering the decision maker perspective, I think</p>	

CHARGE QUESTION 1.a: Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.		
REVIEWER	COMMENT	RESPONSE
	<p>it is important to consider separately the two dimensions: 1) Total number of cases including outbreaks and sporadic cases, and 2) the number of outbreaks. For two food-hazard combinations with the same total number of cases, the one with outbreaks may be considered at highest risk. In the current report sporadic cases are considered in C7 scoring and are not included in C1 scoring. This choice is not logical, in my opinion, because it is done only for hazards for which outbreaks are observed or expected. My proposal is not to combine the two sub-criteria frequency of outbreaks and occurrence of illnesses and have two separate criteria: total number of cases and frequency of outbreaks.</p> <p>Criterion 2: Again as in criterion 1, the interpretation of criterion 2 is not measuring the same things for all the hazards. The hospitalization rate and mortality rate can be used for both chronic and acute exposure. For example, if a chronic exposure to a carcinogenic chemical substance the known type of cancer will inform on the rate of hospitalization (almost 100%) and mortality rate.</p> <p>Criterion 3: Likelihood of contamination. In general, in MCDA, the criteria are independent. Criterion 3 is directly linked to criterion 1. The more likelihood of food-hazard combination contamination is, the higher frequency of illness occurrence might be observed. Criterion 1 is a result of exposure to a particular hazard through the consumption of a particular food. The correlation between these two criteria needs to be addressed when combining the different criteria.</p> <p>Criterion 4: Growth potential/shelf life. Same comment as for criterion 3. There is a possible correlation between criterion 4 and criterion 1.</p> <p>Criterion 5: Manufacturing process contamination probability/intervention. This criterion is in principle correlated to criterion 3. When looking to the attributed scores, it was surprising to find food-hazard combinations scored 0 for criterion 3 being scored 5 for criterion 9 (11 food-hazard out of 1286 food-hazard).</p> <p>Criterion 6: Consumption. The percentage of population consuming food will not capture the entire consumption pattern. Information about frequency of consumption may contribute to the final risk. It will be interesting to know if the food is consumed daily, weekly, monthly, etc.</p> <p>Criterion 7: Economic impact. This criterion should be applied to each hazard and not to each food-hazard combination. To avoid counting two times the total number of cases, it would be better expressed as the average economic impact per case of illness.</p>	

CHARGE QUESTION 1.a: Are the seven criteria used in the draft model appropriate for a multicriteria decision approach? If not, please explain what changes might be considered and why.		
REVIEWER	COMMENT	RESPONSE
Reviewer #3	See response under 1.b.	
Reviewer #4	See response under 1.b.	
Reviewer #5	[To be completed]	

CHARGE QUESTION 1.b: Within the bounds of the FSMA-mandated factors, are there additional criteria beyond the seven criteria that should be considered? If so, please describe these additional criteria that might be considered and why.		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	The seven criteria specified in the draft model appear to cover the FSMA factors – see above.	
Reviewer #2	Figure 2-1 is an effort to explain the relationship between the seven criteria and the FSMA-mandated factors; however factor (v) is not covered. My interpretation of (v), “the likelihood that consuming a particular food will result in a foodborne illness due to contamination of the food”, is different from the one proposed in the current report. In my opinion, factor (v) is not only about exposure but also includes the dose-response relationship. That is, factor (v) tries to capture the infectivity or the toxicity of the hazard. In addition, factor (ii) is not totally covered because criterion 2 is not considering the level of contamination for chemical hazards, unless the frequency of contamination is taking into account only events with high level of contamination.	
Reviewer #3	The seven criteria selected by FDA seem to be reasonable and well-justified based on the body of available scientific knowledge, but detailed knowledge of FSMA is outside of my area of expertise.	
Reviewer #4	For convenience and clarity I have combined the responses to the above two questions. Food safety risk is normally interpreted as the number and severity of <i>potential</i> or actual cases of foodborne illness related to some defined food(s) and hazard(s). To assess such risk it is necessary to consider the probability that a food will be contaminated with the hazard of interest, the number of people who eat that food and how frequently, the probability that the level of contamination will be high enough to cause illness, and severity of the resultant illness. “Severity” encompasses the severity of disease symptoms and duration of illness experienced by a consumer (often evaluated using the DALY or QALY metrics) but may also consider the <i>absolute</i> number of cases. For microbial hazards, the probability of contamination at a level likely to cause illness is also related to the potential for growth of the organism in the food. The potential for growth is a function of the specific ecophysiology of the hazardous organism, the composition and packaging of the food (esp. gaseous atmosphere) and the time, temperature,	

CHARGE QUESTION 1.b: <i>Within the bounds of the FSMA-mandated factors, are there additional criteria beyond the seven criteria that should be considered? If so, please describe these additional criteria that might be considered and why.</i>		
REVIEWER	COMMENT	RESPONSE
	<p>(and to a lesser extent) the relative humidity of the storage environment. In the Risk Ranking Model economic ‘impact’ was also required to be included as part of the overall risk, though the economic impact assessment was limited to health care costs, and did not include costs to industry from an outbreak, etc.</p> <p>The factors nominated by FMSA for inclusion in the designation of HRF foods were:</p> <ul style="list-style-type: none"> i) the known safety risks of a particular food, including the history and severity of foodborne illness outbreaks attributed to such food, taking into consideration foodborne illness data collected by the Centers for Disease Control and Prevention (CDC); ii) the likelihood that a particular food has a high potential risk for microbiological or chemical contamination or would support the growth of pathogenic microorganisms due to the nature of the food or the processes used to produce such food; iii) the point in the manufacturing process of the food where contamination is most likely to occur; iv) the likelihood of contamination and steps taken during the manufacturing process to reduce the possibility of contamination; v) the likelihood that consuming a particular food will result in a foodborne illness due to contamination of the food; and vi) the likely or known severity, including health and economic impacts, of a foodborne illness attributed to a particular food. <p>These nominated factors include consideration of factors generally considered to affect foodborne health risk, though in some cases in a somewhat cryptic way, e.g., the known safety risks from epidemiological data implicitly encompass many of the other factors that the more inductive approach to risk assessment considers explicitly. This potentially introduces a confounding in the approach to the risk ranking because some factors will, in effect, be considered twice. (This aspect was discussed in the complementary data review report).</p> <p>In the above set of factors, however, the total consumption is not explicitly nominated as a risk-affecting factor, except that it is implicit in the epidemiological data because, all other things being equal, foods that are more frequently consumed are more likely to be involved in detectable</p>	

CHARGE QUESTION 1.b: <i>Within the bounds of the FSMA-mandated factors, are there additional criteria beyond the seven criteria that should be considered? If so, please describe these additional criteria that might be considered and why.</i>		
REVIEWER	COMMENT	RESPONSE
	<p>outbreaks. This is not necessarily an oversight in the identification of risk-affecting factors. In managing risk, or in identification of a “HRF”, it may be the <i>intrinsic</i> properties of a food and the processing, packaging and storage conditions that are considered to be the most important components for risk, i.e., to be able to estimate the <i>risk-per-serving</i>.</p> <p>In the development of the risk ranking model the six factors above were translated into seven criteria:</p> <ul style="list-style-type: none"> Criterion 1. Frequency of outbreaks and occurrence of illnesses (i.e., Epidemiological Link); Criterion 2. Severity of illness, taking into account illness duration, hospitalization and mortality Criterion 3. Likelihood of contamination Criterion 4. Growth potential/shelf life Criterion 5. Manufacturing process contamination probability/intervention Criterion 6. Consumption Criterion 7. Economic impact <p>The mapping of these seven criteria to FSMA’s six factors is shown diagrammatically in Figure 2-1 in the Draft Report, which clearly shows the some of the criteria are confounded, e.g., as suggested above, the epidemiological data would be expected to be a reflection of other risk-influencing factors (Criteria 3-6), so that effectively a “double-counting” of some factors is occurring in the logic of the model. Notably, in the FDA model, consumption is explicitly added to the risk ranking criteria, despite that it appears not to have been explicitly considered by FMSA. As such, there may need to be clarification of whether the intent of the RRM is to rank risk on a per-serving basis, or on a whole of production/supply basis.</p> <p>In the draft model, however, one necessary risk-affecting factor that appears to be inadequately addressed is the relevance of the <i>level</i> of contamination, i.e., consideration of the dose-response relationship and its influence on the inferred probability of illness. This issue is implicit in FMSA factor v). It is also apparent from Figure 2-1 that it is <i>not</i> explicitly considered in the draft risk ranking model, i.e., because factor v) is ‘covered’ only by consumption data. The lack of an</p>	

CHARGE QUESTION 1.b: <i>Within the bounds of the FSMA-mandated factors, are there additional criteria beyond the seven criteria that should be considered? If so, please describe these additional criteria that might be considered and why.</i>		
REVIEWER	COMMENT	RESPONSE
	<p>infectious/toxic dose consideration seems curious given that the draft model is stated to have been based on FDA's fresh produce risk ranking tool which <i>does</i> consider 'infectious dose'.</p> <p>Another consideration that is, perhaps, implicitly suggested in FSMA criterion iii), but not explicitly considered in the FDA draft risk model, is the influence of the <i>sequence</i> of risk affecting influences. An obvious example is a heating step for a food in a sealed container, or a cooking step prior to consumption, that eliminates a microbial risk completely. Conversely, if recontamination and growth is possible after a 'kill' step, the risk will be completely different. The draft model does not reflect the significance of such differences in the <i>sequence</i> of steps and cannot represent this using the additive scoring process adopted, as will be illustrated more clearly in responses to later questions.</p>	
Reviewer #5	[To be completed]	

CHARGE QUESTION 2: <i>Are the scoring definitions for all criteria appropriate?</i>		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	The scoring definitions for the seven criteria are generally appropriate for microbial risks and undeclared allergens. Some scoring definitions for chemicals are difficult to follow and commented further below.	
Reviewer #2	I will say that the definitions are in general clear and the provided document allows for the needed verifications. Please see 2b.	
Reviewer #3	See response under 2.a.	
Reviewer #4	See response under 2.a.	
Reviewer #5	[To be completed]	

CHARGE QUESTION 2.a: <i>Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and undeclared allergens)? If not, please describe changes that might be considered and why.</i>		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	<p><i>Criterion 1 (frequency of outbreaks and occurrence of illness)</i></p> <p>For chemicals, it is said that scoring is based on expert elicitations and the definition for the scores (1, 3, and 9) uses terms "little, some, and compelling" evidence. What constitute little, some and compelling? Was there a rigorous evidence-based approach with guiding principal that was applied for consistency in</p>	

CHARGE QUESTION 2.a: Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and undeclared allergens)? If not, please describe changes that might be considered and why.		
REVIEWER	COMMENT	RESPONSE
	<p>reaching these ratings? Further by examining the list of experts from which the scores were elicited, there are 3 experts labelled as “toxicologists”. Given the range of chemicals potentially involved, such elicitation process would benefit from a larger pool of toxicologists.</p> <p>Per appendix K, Table 5, there are 160 cases where expert elicitation was used to assign scores to criterion 1 for chemicals, 33 for undeclared allergens, and none for microbes. Given the emphasis of the expert opinion on the value/score of criterion 1 for chemical, a more robust elicitation (i.e., larger pool of experts) to capture range of expert opinions would be warranted.</p> <p>Also for undeclared allergens, appendix K, page 7, noted that scores were elicited from an allergens expert separately from the microbial and chemical hazard groups. However, from the list of experts (Table 4-3 in draft report, or Table 2 of appendix K), there are no allergen experts. For transparency, the allergen experts should be provided in the draft report.</p> <p><i>Criterion 2 (severity of illness)</i> For chemicals it is said that the definition for scoring severity with acute exposure is based out of ICMSF (2001) and used for scoring in the draft risk model, according to the definitions in Table 2.1. Similarly, for chronic the definitions in Table 2.2 are applied for scoring. Further into Section 4.2.2 of the draft report, it appears that the scoring was done by subject matter experts using the definitions in 2.1 and 2.2. Who are these subject matter experts? Was this done through the same expert elicitation process as described in Appendix K? Per appendix K, there are 25 cases where expert elicitation was used but there are more than 25 cases of food-chemical pairs in the draft model. More transparency is needed in derivation of the scores for criterion 2 for chemicals.</p> <p><i>Criterion 3 (likelihood of contamination)</i> For chemical hazards, this likelihood of contamination is said to be determined based on percent positive above action levels or allowable levels. Are the weights (n*gw*dw) that is applicable in the case of microbes also applicable to chemicals (and allergens)? They should be if they are not already.</p> <p>The USDA Agricultural Marketing Service (AMS) Pesticide Data Program (PDP) is a potential data source to rely upon for scoring this criterion. Is there a reason why it was not included as a reference source?</p>	

CHARGE QUESTION 2.a: Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and undeclared allergens)? If not, please describe changes that might be considered and why.

REVIEWER	COMMENT	RESPONSE																							
	<p><i>Criterion 4 (growth potential /shelf life)</i> This criterion is strictly for microbes. The current method of aggregation of scores across seven criteria to derive the composite score for each food-hazard pair by summing (equal weights to all seven criteria) is preferentially selecting food-microbe pairs (i.e., forcing higher ranks on these pairs over chemical and undeclared allergens where there is usually no growth).</p> <p>To avoid the current imbalance, both criteria 3 and 4 are relevant to address FSMA factor ii (i.e., “the likelihood that a particular food has a high potential risk for microbiological or chemical contamination <i>or</i> would support the growth of pathogenic microorganisms due to the nature of the food or the processes used to produce such food”. For chemical, criterion 3 is relevant to address FSMA ii. For microbial hazard, conceptually the score for FSMA ii can be derived based on the composite of criteria 3 and 4 as followed:</p> <table border="1" data-bbox="506 753 1505 959"> <tr> <td rowspan="3">C3 Likelihood of contamination</td> <td>High (9)</td> <td>3</td> <td>9</td> <td>9</td> </tr> <tr> <td>Medium (3)</td> <td>1</td> <td>3</td> <td>9</td> </tr> <tr> <td>Low (1)</td> <td>1</td> <td>1</td> <td>3</td> </tr> <tr> <td></td> <td></td> <td>Low (1)</td> <td>Medium (3)</td> <td>High (9)</td> </tr> <tr> <td></td> <td></td> <td colspan="3">C4 Likelihood of growth</td> </tr> </table> <p><i>Criterion 6 (consumption)</i> In the current draft model/report, it is stated that the consideration of both criteria 3 (likelihood of contamination) and C6 (percent consumer) defines the likelihood of consuming a particular contaminated food result in illness. While the percent consumer may be a reasonable proxy for microbial and undeclared allergen risk, for chemical risk, the dose make the poison. Thus, there is a need to know how much is consumed, i.e. the likelihood of consuming a particular contaminated food resulting in illness is a consideration of both criterion 3 (likelihood of contamination) and criterion 6 (percent consumer*amount consumed). Conceptually, scoring for criterion 6 for chemicals may be as followed:</p>	C3 Likelihood of contamination	High (9)	3	9	9	Medium (3)	1	3	9	Low (1)	1	1	3			Low (1)	Medium (3)	High (9)			C4 Likelihood of growth			
C3 Likelihood of contamination	High (9)		3	9	9																				
	Medium (3)		1	3	9																				
	Low (1)	1	1	3																					
		Low (1)	Medium (3)	High (9)																					
		C4 Likelihood of growth																							

CHARGE QUESTION 2.a: Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and undeclared allergens)? If not, please describe changes that might be considered and why.

REVIEWER	COMMENT	RESPONSE																							
	<table border="1" data-bbox="600 418 1505 594"> <tr> <td rowspan="3">Percent consumers</td> <td>High</td> <td>3</td> <td>9</td> <td>9</td> </tr> <tr> <td>Medium</td> <td>1</td> <td>3</td> <td>9</td> </tr> <tr> <td>Low</td> <td>1</td> <td>1</td> <td>3</td> </tr> <tr> <td></td> <td></td> <td>Low</td> <td>Medium</td> <td>Day</td> </tr> <tr> <td></td> <td></td> <td colspan="3">Amount consumed per day (g/day)</td> </tr> </table> <p data-bbox="394 630 1646 930"><i>Criteria 7 – economic impact</i> Scallan et al. (2011) and Palmer et al. (2013) are noted as sources for this information for scoring outbreaks and sporadic cases (page 24 of draft report). What are the reference sources for chemical related endpoints? What are the endpoints that were captured in this metric? Was expert elicitation utilized for missing data for chemical endpoints? If so, were health economists among the experts from which the information was elicited? On page 66 of the draft report, Figure 4-7 indicates that expert judgment is used when no quantitative data are available. Who are these experts? Transparency is needed here. In the example on apple juice-arsenic, page 68, how are the dollar amount assigned to the 51.4 cancer cases/year?</p>	Percent consumers	High	3	9	9	Medium	1	3	9	Low	1	1	3			Low	Medium	Day			Amount consumed per day (g/day)			
Percent consumers	High		3	9	9																				
	Medium		1	3	9																				
	Low	1	1	3																					
		Low	Medium	Day																					
		Amount consumed per day (g/day)																							
Reviewer #2	No comment.																								
Reviewer #3	Various types of hazards appear to be defined well, but detailed hazard identification for foods is outside of my area of expertise.																								
Reviewer #4	<p data-bbox="394 1044 1646 1307">The hazards are appropriately described for microbial, chemical and allergen hazards. Some of the chemical hazards listed are of microbial origin (algal and fungal toxins), but are usually already present in the food (prior to processing) and would not increase in level during processing, distribution etc. Histamine levels can increase as microbial metabolism continues, and this difference is appropriately identified in the approach developed, and dealt with by considering histamine as a microbial hazard. Nonetheless, fungal growth and aflatoxin production is also possible in many foods after processing, and the difference in approach for histamine, compared to aflatoxin, requires further explanation for transparency of the document and approach.</p> <p data-bbox="394 1344 1646 1404">The relativities of some scores do not seem correct. For example, hazard severity scores for microbial hazards vs. allergens do not accord with independent expert assessment by Minor <i>et al.</i> 2015) and it was</p>																								

CHARGE QUESTION 2.a: Are the definitions appropriately defined for the various types of hazards considered (i.e., microbial hazards, chemical hazards (including chronic exposure) and undeclared allergens)? If not, please describe changes that might be considered and why.		
REVIEWER	COMMENT	RESPONSE
	notable that 70% of the highest ranked hazards for all food-hazard pairs were for allergens, based only on expert opinion rather than data. This does not seem to accord with the available public health data. There is more detailed discussion of these anomalies in the complementary data review report.	
Reviewer #5	[To be completed]	

CHARGE QUESTION 2.b: Is the value function 0-1-3-9 and scoring matrix appropriate for the intended purpose to inform the designation of high-risk foods? If not, please describe changes that might be considered and why.		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	<p>The sensitive analysis in Appendix N (Section 6. Impact of scoring scale) explored the use of an alternative ordinal scale of 1, 2, 3 and 4 (i.e., Anderson model) and impact on model results. The output in Table 9 of the Appendix N showed somewhat comparable results for the top 20 commodities when using the 0, 1, 3, and 9 (the current model) and when using the Anderson models, and it was concluded that the scoring scale in the HRF model did not dramatically affect the ranking results. The following suggestions are made to expand the sensitivity analysis to strengthen the current conclusion that ranking of foods is insensitive to the types of value function being used:</p> <ul style="list-style-type: none"> • Since there are 335 specific foods in the current model, the table showing the comparability of ranks should be expanded to show comparability of the top 10-20% ranks from each type of ranking. <p>The comparability of the value function being used to score the seven criteria should also be examined at the food-hazard pair level. Insight on whether a value function has an influence on the rank order of the food-hazard pairs, can help inform their potential impact on the aggregated scores for a food, depending on what method of aggregation across multiple hazards.</p>	
Reviewer #2	<p>Observed data and information related to the different criteria are grouped into scoring bins, which are defined and assigned a numerical value from 0 to 9. The numerical values from 0 to 9 reflect categories and their assignment is arbitrary, except that the values reflect the increasing quantity of the criteria being measured. The scale of the scoring is ordinal despite the distance between the possible values 0-1-3-9. The scoring 0-1-3-9 cannot be interpreted as an interval scale or as a ratio scale. Adding ordinal scales cannot be done in principle. Adding ordinal scales may lead to incorrect conclusions and this is the main challenge for risk ranking based on ordinal scales and the motivation for developing adequate combination of criteria measured with ordinal scales. The proposed ordinal scales intentionally leave gaps between the</p>	

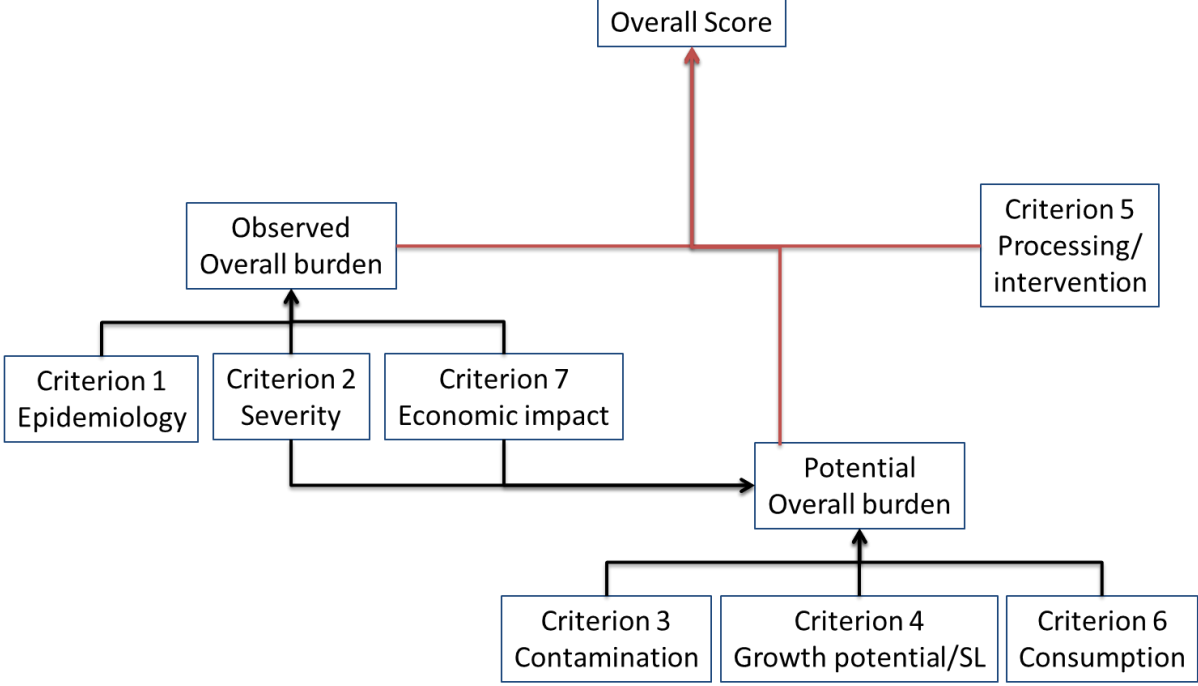
CHARGE QUESTION 2.b: Is the value function 0-1-3-9 and scoring matrix appropriate for the intended purpose to inform the designation of high-risk foods? If not, please describe changes that might be considered and why.		
REVIEWER	COMMENT	RESPONSE
	<p>numerical values to better represent the assumed distance between categories is a possible approach but does not solve the entire problem.</p> <p>The gaps between the numerical values are not consistent for all the criteria. For example, the significance of zero is for criteria absence of event and for other absence of data. The corresponding bins for 1-3-9 for one criteria are defined using a linear scale (i.e., consumption: 0 (>1%), 1 (1-5%), 3 (5-10%), 9(>10%)) and for another a sort of log-scale (i.e., contamination: 0 (No known occurrence), 1(<0.1%), 3(0.1-1%), 9(>1%)).</p>	
Reviewer #3	<p>It is unclear whether the proposed value function and scoring matrix are appropriate to inform the designation of high-risk foods. On one hand, it is certainly true that for particular assignments of scores, a higher score is associated with a higher risk. On the other, it is unclear whether the scores are meant to indicate the magnitude of a risk factor, its order-of-magnitude of a risk, or something else entirely. For example, is a 9 actually supposed to be three times as bad as a 3? In the example on p. 17, a hazard with a hospitalization rate of 9% would have a score of 1, and a hazard with a hospitalization rate of 21% would have a score of 9.</p> <p>This only becomes more complicated and unclear for the case of non-numerical constructed scales. In a sense, the point of these scales is to serve as proxies for probabilities whose explicit elicitation will be impractical. It would help if the scales were constructed with consistent intent, e.g., with the intent that scores correlate with the magnitude of the risk. As will be discussed in my later comments, because these scores are to be used in further calculations, it is important to be clear on what they are supposed to represent, since their properties determine when it is appropriate to perform such calculations and how to interpret their results. For example, statistics courses teach that averages of ordinal data are not meaningful – are three satisfied customers the same as one dissatisfied customer, one satisfied customer and one extremely satisfied customer?</p>	
Reviewer #4	<p>Additive scoring schemes, even if based on exponential scales, cannot capture the full complexity of factors and their interactions that contribute to food-borne risk.</p> <p>All scoring schemes that attempt to translate continuous variables affecting risk into a limited number of categories and assign discrete numerical values will distort the calculations and relativities of risk. This compromise is sometimes necessary or useful when having to do many sets of calculations with imperfect data and for diverse properties. The sort of semi-quantitative scoring scheme adopted here is useful when one does not have access to a computer to do the complex calculations involved in quantitative risk assessment. However, given that the system described here is for government use, and is already presented</p>	

CHARGE QUESTION 2.b: <i>Is the value function 0-1-3-9 and scoring matrix appropriate for the intended purpose to inform the designation of high-risk foods? If not, please describe changes that might be considered and why.</i>		
REVIEWER	COMMENT	RESPONSE
	<p>as relatively sophisticated computer software, the reasons for adopting what is essentially a complex decision matrix rather than a generic quantitative risk assessment model, e.g., <i>iRisk</i>, are unclear and seem to require further explanation in the draft RRM report and accompanying documents and tools. The relationships between risk-influencing factors and how they contribute to risk can be expressed algebraically so that the full value of numerical data and the understanding of the interactions of factors that lead to risk could be preserved. It may not have been possible in this risk ranking project, e.g. due to time or personnel constraints, to collate and analyze all the data for each food-hazard pair into a representative value and, in those circumstances, a semi-quantitative approach as adopted may make sense. However, further justification/explanation of the adoption of a semi-quantitative risk matrix approach would be useful and contribute to the transparency of the approach and model.</p> <p><i>Prima facie</i>, the use of a quasi-exponential approach in scoring (i.e., where a difference of three in a score typically represents a factor of ten difference in risk, or a score difference of 9 represents a thousand-fold difference in risk) also makes sense to reflect that most (though not all) food safety risk-affecting risk factors multiply together to generate the risk. However, microbial risk assessments, in particular, often involve calculations with, and combination of, factors that operate naturally on exponential scales and others that operate naturally on arithmetic scales. This can lead to problems in semi-quantitative assessments based on discrete categories, as illustrated below.</p> <p>Discretization of continuous variables into a small number of ‘representative’ values, or categories, can be expected to produce anomalies. Consider, for example, two product:hazard pairs. Let hazard ‘A’ be at the lower end of the ‘middle’ range for three factors and hazard ‘B’ be at the upper end of ‘middle’ range for the same three factors. Each range represents roughly a factor of 10. Thus, while both hazards would score the same (i.e., $3+3+3 = 9$), the true difference in risk would be expected to be closer to 1000-fold, i.e., hazard ‘B’ should have a risk score 9 higher than hazard ‘A’. As a second example, imagine that hazard ‘A’ is at the upper end of the ‘middle’ range for 3 factors, while hazard ‘B’ is at the lower end of the ‘high’ range for 3 factors. Now hazard ‘A’ has a score of 9, while hazard ‘B’ has a score of 27. On the basis of the quasi-exponential scale, the risk from hazard ‘B’ is now evaluated to be 1 billion times greater than the risk from hazard ‘A’ even though we know they are only subtly different in their levels in each of the three categories.</p> <p>As a further example, consider three pairs of food:hazards. Hazard 1 has three attributes at the upper end of the ‘low’ range, and has a risk score of 3 ($1+1+1$). Hazard 2 has three attributes in the middle on of the ‘intermediate’ range and receives a score of 9 ($3+3+3$). Implicitly, Hazard 2 is 1000-fold more risky than</p>	

CHARGE QUESTION 2.b: <i>Is the value function 0-1-3-9 and scoring matrix appropriate for the intended purpose to inform the designation of high-risk foods? If not, please describe changes that might be considered and why.</i>		
REVIEWER	COMMENT	RESPONSE
	<p>Hazard 1, yet the score difference is 6 (implicitly only two factors of 10). Now, consider Hazard 3 which has three attributes at the low end of the relevant ‘high’ range. It would be scored 27 but, implicitly Hazard 3 is ~1000-fold more risky than Hazard 2. The score difference, however, is 18 suggesting 6 orders of magnitude difference, rather than the 3 orders ‘built into’ this scenario. The relative risk from Hazard 2 compared to Hazard 1 is the same as the relative risk from Hazard 3 to Hazard 2, yet the score differences are completely different. Thus, when used in scoring the overall risk, the relative risk scales inherent in the scoring for individual criteria become distorted, and the final risk score seems more like an absolute scale rather than an exponential scale. The combination and interconversion of factors that operate on arithmetic <i>cf.</i> logarithmic scales is a recognized source of logical errors in food safety risk assessment particularly for risks from microbial hazards.</p> <p>As another example (mentioned earlier) of logical weakness in the scoring scheme, a listericidal process in a hermetically sealed product effectively eliminates the risk from that hazard in that product, irrespective of what occurred before or what will happen later to that product, as long as package integrity is preserved. Equally, proper cooking can completely eliminate many hazards. The current scoring scheme does not enable a “reset to zero” for ‘cidal’ processes and does not correctly represent such scenarios.</p> <p>These examples are not presented to suggest that the model requires change, but that <i>users will have to be very careful about the interpretation of the risk ranking results</i>, particularly if one is trying to assess relative risk, <i>and particularly if trying to establish a threshold risk value that defines an HRF.</i></p> <p>Given the above discussion, however, there does appear to be a logical deficiency in the RRM and it is suggested that the approach/considerations used to define FDA’s ‘potentially hazardous foods’ or ‘temperature controlled for safety’ foods (<i>see also</i> response to Q.6) may offer a good starting point, or useful insights, for the proposed risk ranking model despite that it will only offer insights about ranking of microbial hazards. It also would seem to offer the advantage of policy consistency.</p> <p>As discussed more fully in the complementary report on the model data, while no simple scoring scheme is likely to correctly rank all product:hazard pairs and will of necessity be a compromise, it may be possible, nonetheless, to achieve consensus among stakeholders, i.e., to accept the RRM as being the best approach possible, or at least as good as any other approach, if stakeholders agree that the RRM does produce a ranking that ‘make sense’. Users should remain aware, however, of those model limitations and potential anomalies and not accept the model results without some form of scrutiny, or ‘reality checks’.</p>	
Reviewer #5	[To be completed]	

CHARGE QUESTION 3: <i>Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.</i>		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	In the current draft model, the algorithm to generate a risk score for a particular food-hazard pair (FRS) is the sum of the weighted scores for the seven criteria. What was the rationale for addition as the mathematical operation to combine the seven criteria to derive a composite score of each food-hazard pair? Were other options to combine considered and sensitive analyses done? Other options may include weighted average of the seven criteria, weighted product of the seven criteria, or a combination of weighted addition of likelihood for exposure indicators and likelihood of illness indicators and multiplication of these two composite likelihoods (i.e., likelihood of risk = exposure dose x dose-response). At the very least, this should be discussed and rationale provided as to why they were not pursued.	
Reviewer #2	<p><u>Aggregation</u></p> <p>Considering the nature of the measurement scales, I think that the proposed algorithm that combines criteria scores and weights is not appropriate.</p> <p>Before presenting the suggestions to improve the current algorithm, it is important to know exactly the significance of the overall score. Because of the possible non-independency between the seven criteria, there is a need to create a sort of hierarchy between the seven criteria. One of the possible restructurings is as follow:</p>	

CHARGE QUESTION 3: Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.

REVIEWER	COMMENT	RESPONSE
	 <p>The flowchart illustrates the risk ranking structure. At the top is 'Overall Score'. Below it are 'Observed Overall burden' and 'Potential Overall burden'. 'Observed Overall burden' is influenced by 'Criterion 1 Epidemiology', 'Criterion 2 Severity', and 'Criterion 7 Economic impact'. 'Potential Overall burden' is influenced by 'Criterion 3 Contamination', 'Criterion 4 Growth potential/SL', and 'Criterion 6 Consumption'. Additionally, 'Criterion 5 Processing/intervention' is shown as a separate factor that influences the 'Overall Score'.</p> <p>Figure 1: Risk ranking structure. Criterion 1 includes sporadic cases, criterion 7 economic impact per case of illness.</p> <p>The assumptions of the new structure are:</p> <ul style="list-style-type: none"> - A food-hazard combination is assumed to be at high risk if the observed overall burden is high, or the potential overall burden is high or the criterion 5 (probability of contamination at processing is high and weak intervention). - The observed overall burden is an assessment of the risk based on empirical epidemiological evidence: “top down” assessment - The potential overall burden is an assessment of the risk based on empirical food chain evidence (contamination, consumption, growth...): “bottom up” assessment. <p>Two rankings may be performed: one considering “Observed overall burden” and the other “Potential overall burden”.</p>	

CHARGE QUESTION 3: *Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.*

REVIEWER	COMMENT	RESPONSE
	<p>Using the report structuring of the criteria or the one proposed in Figure 1, the combination of the score attributed to the different criteria need to be modified:</p> <p>As the scores are assigned using ordinal scales, it is not possible to use the weighted sum to aggregate the seven criteria. One of the possible solutions is to use the outranking concepts (Roy, 1978). An outranking relation of two food-hazard combinations a and b, as a binary relation S defined on a set of food-hazard combinations A, such that aSb (a outranks b) if, given what is known about the decision maker's preferences, and given the evaluations on food-hazard combinations and the nature of the problem, there exist enough arguments to decide that a is at least as risky as b, while there is no essential reason to disapprove that statement. To implement the outranking concept, one can use one of the ELECTRE methods (Elimination and Choice Translating Reality, ELECTRE I, II, III, IV). ELECTRE III method was designed to deal with inaccurate, imprecise or uncertain data. This method utilize pseudo-criteria instead of the defined and is suitable if at least one of the following situations is shown:</p> <ul style="list-style-type: none"> - Ranking problem where the ranked objects are evaluated (for at least one criterion) on an ordinal scale or on a weak interval scale. These scales are not suitable for the comparison of differences. - A strong heterogeneity related to the ways criteria are evaluated which makes it difficult to aggregate all the criteria in a unique and a common scale - Compensation of the loss on a given criterion by a gain on another one may not be acceptable for the decision maker. Then, such situations require the use of no compensatory aggregation procedures. - For at least one criterion small differences are not significant in terms of preferences, while the accumulation of several small differences may become significant. This requires the introduction of discrimination thresholds. <p>The main purpose of ELECTRE III method is to rank the food-hazard combinations based on two indices, the concordance index and the discordance index defined for each pair of food-hazard combinations a and b.</p> <p>The concordance index c(a,b) is calculated by:</p> $c(a, b) = \frac{1}{W} \sum_j w_j c_j(a, b)$ <p>where j is one of the seven criteria, W is the sum of weights of the different criteria (wj) and</p>	

CHARGE QUESTION 3: Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.

REVIEWER	COMMENT	RESPONSE
	$c_j(a, b) = \begin{cases} 1 & \text{if } sj(a) - sj(b) < q_j \\ 0 & \text{if } q_j < sj(a) - sj(b) < p_j \\ \frac{p_j - (sj(a) - sj(b))}{p_j - q_j} & \text{if } sj(a) - sj(b) > p_j \end{cases}$ <p>sj(a) is the score for criterion j for food-hazard combination a, pi and qj are discrimination thresholds that define zones of strict difference, indifference and weak difference. The concordance index evaluate to what extent (a) is at least as high risk as (b).</p> $d_j(a, b) = \begin{cases} 1 & \text{if } sj(a) - sj(b) > v_j \\ 0 & \text{if } p_j < sj(a) - sj(b) < v_j \\ \frac{p_j - (sj(a) - sj(b))}{p_j - q_j} & \text{if } sj(a) - sj(b) < p_j \end{cases}$ <p>Where vj is the veto threshold.</p> <p>The overall concordance and discordance indices are then combined to obtain a valued outranking relation with credibility that a outranks b defined as:</p> $\delta(a, b) = \begin{cases} c(a, b) & \text{if all } dj(a, b) \leq c(a, b) \\ c(a, b) \prod_{i \in J(a,b)} \frac{1 - dj(a, b)}{1 - c(a, b)} & \end{cases}$ <p>Where J(a,b) is the set of criteria j for which dj(a,b)>c(a,b). The discordance Index to what extent the overall difference between the scores of (a) and (b) is enough important that (a) is not as high risk as (b).</p> <p>The overall concordance and discordance indices are than used to provide two rankings: descending and ascending ranking. And the combination of the two ranking provides the final ranking.</p>	

CHARGE QUESTION 3: *Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.*

REVIEWER	COMMENT	RESPONSE
	<p>The implementation of ELECTRE methods can be done using available software (http://www.lamsade.dauphine.fr/english/software.html).</p> <p>References ELECTRE III:</p> <ul style="list-style-type: none"> - Rainer Bruggemann1 and Lars Carlsen. Incomparable - What Now, III. Incomparabilities, Elucidated by a Simple Version of ELECTRE III and a Fuzzy Partial Order Approach. MATCH Commun. Math. Comput. Chem. 73 (2015) 277-302 - José Figueira, Salvatore Greco, Bernard Roy and Roman Slowinski. Electre Methods: Main Features and Recent Developments. HAL Id: hal-00876980 https://hal.archives-ouvertes.fr/hal-00876980 25 Oct 2013. - <p><u>Weights</u></p> <p>The approach used to elicit the possible weight is not well formalized.</p> <p>First proposal: organize an expert elicitation using for example AHP method.</p> <p>Second proposal: use constraint algorithm (sum of weight equal to 1) and learning system. The learning system can be organized by providing decision makers/stakeholders a random set of valuated criteria and ask the participants to rank them without providing any explanation about their outcomes.</p>	
Reviewer #3	<p>The overall approach is reasonable. In a perfect world, an ideal model (such as a full, vetted, detailed risk assessment model) would calculate explicit risks associated with different foods and compare them to benchmark levels derived using solid toxicology models. However, with various foods to consider and limited information to analyze in this vein, such a model is practically difficult to adequately develop. At the other extreme, a simple model might just require a food expert or panel of experts to give a qualitative score to each food and rank them accordingly. This proposed model lies in between, providing some structured knowledge base: the data fields used, the algorithm for calculating risk scores, and the weighting parameters and scoring protocols used are intended to get more of the precision and transparency of the ideal model without creating onerous information requirements. To the extent that it does this in a way that approximates what an ideal model would do, the approach can lead to real practical improvement in food prioritization by taking the best course of action to overcome current knowledge limitations.</p> <p>With this said, I have concerns about specific algorithms used. It seems the factors should be multiplicative because they form a chain of events leading to consequences. Ideally, the overall score would be highest for the risks with the greatest average consequences, i.e., foods that are classified as high-risk would have higher risk than foods that are classified as low risk. This may not happen if the hazard scores represent magnitudes, or are approximately linear in probability. For example, hazard A</p>	

CHARGE QUESTION 3: <i>Is the algorithm that combines criteria scores and weights into an overall score appropriate? If not, please provide suggestions on what improvements should be considered.</i>		
REVIEWER	COMMENT	RESPONSE
	<p>with a 50% chance of causing 5 units of harm has an expected loss of 2.5 units, while hazard B with a 90% chance of causing 2 units of harm has an expected loss of 1.8 units. However, if the factor scores are translated to (5, 5) for hazard A and (9, 2) for hazard B, hazard A has a lower aggregate risk score (10) than hazard B (11).</p> <p>If created and applied with care, factor scores that reflect order-of-magnitude of risk (and that also represent sequential events) can be added in such a way that the higher score is associated with higher risk. This is due to the fact that adding such factor scores would be equivalent to multiplying the magnitude of risks at each stage. For aggregation within a food group, a mathematically sound approach along these lines (that proxies for risk analysis) is even harder to figure out. The problem is, there are no easy shortcuts to producing a score that has the precision of a structured risk analysis. This is better than nothing if the scores are handled as coarse indicators. Their numerical nature should not be confused with across-the-board rigor.</p> <p>More details are presented in the additional comments at the end of this review.</p>	
Reviewer #4	<p>See comments above (in response to Q2b) which comments on and exemplifies potential problem in the scoring scheme rather than the “algorithm”. An alternative approach was suggested there. The calculations, including selection of weighting factors are very simple and straightforward, i.e., simple additions of the seven criteria scores after application of any weighting. (In fact, the use of the term algorithm is not really relevant in this case because the calculation can be expressed as a single equation). Importantly the calculations are transparently presented and explained in the Draft Report with one possible exception (<i>see</i> response to Question 7b).</p>	
Reviewer #5	[To be completed]	

CHARGE QUESTION 4: <i>Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.</i>		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	<p>In the current draft model, the default weight of 10 is applied to all seven criteria, i.e., no differential weighting, and the scores for the seven criteria are summed to derive FRS for each food-hazard pair.</p>	

CHARGE QUESTION 4: *Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.*

REVIEWER	COMMENT	RESPONSE																																																												
	<p>As noted earlier, the current summation with criterion 4 given full weight, would preferentially rank food-micro hazard pairs higher than chemical and undeclared allergen pairs. Suggestion is provided above to avoid the imbalance between the hazard types.</p> <p>Since the intent of the model is to address FSMA factors/requirements and the seven criteria that were derived intentionally to capture the FSMA factors, it may be best to focus the weighting of these seven criteria based on FSMA emphasis. The table below assumed a default equal weight of 10 for each of the six FSMA factors (i.e., FSMA weight), hence a total weight of 60. The table also summarized the seven criteria in the draft model per each FSMA factor and the FSMA weight for each of the criterion is derived by dividing the FSMA default weight by the number of the model criteria assigned to the FSMA factor. The final weight for each criterion is derived by summing the FSMA weight. Criterion 4 is only applicable to micro hazards so noted with C3 (see above for suggestion to composite with C3 for micro hazard).</p> <table border="1" data-bbox="401 824 1642 1243"> <thead> <tr> <th>FSMA Factor</th> <th>FSMA Weight</th> <th>Criteria</th> <th>FSMA weight</th> <th colspan="2">Final criterion weight</th> </tr> </thead> <tbody> <tr> <td rowspan="2">FSMA i</td> <td rowspan="2">10</td> <td>C1</td> <td>5</td> <td>C1</td> <td>5</td> </tr> <tr> <td>C2</td> <td>5</td> <td>C2</td> <td>5</td> </tr> <tr> <td rowspan="2">FSMA ii</td> <td rowspan="2">10</td> <td>C3</td> <td>5</td> <td colspan="2">C3 (&C4) 15</td> </tr> <tr> <td>C4</td> <td>5</td> <td>C5</td> <td>20</td> </tr> <tr> <td rowspan="2">FSMA v</td> <td rowspan="2">10</td> <td>C3</td> <td>5</td> <td>C6</td> <td>5</td> </tr> <tr> <td>C6</td> <td>5</td> <td>C7</td> <td>10</td> </tr> <tr> <td>FSMA iii</td> <td>10</td> <td>C5</td> <td>10</td> <td></td> <td></td> </tr> <tr> <td>FSMA iv</td> <td>10</td> <td>C5</td> <td>10</td> <td></td> <td></td> </tr> <tr> <td>FSMA vi</td> <td>10</td> <td>C7</td> <td>10</td> <td></td> <td></td> </tr> <tr> <td>Total</td> <td>60</td> <td></td> <td>60</td> <td></td> <td></td> </tr> </tbody> </table>	FSMA Factor	FSMA Weight	Criteria	FSMA weight	Final criterion weight		FSMA i	10	C1	5	C1	5	C2	5	C2	5	FSMA ii	10	C3	5	C3 (&C4) 15		C4	5	C5	20	FSMA v	10	C3	5	C6	5	C6	5	C7	10	FSMA iii	10	C5	10			FSMA iv	10	C5	10			FSMA vi	10	C7	10			Total	60		60			
FSMA Factor	FSMA Weight	Criteria	FSMA weight	Final criterion weight																																																										
FSMA i	10	C1	5	C1	5																																																									
		C2	5	C2	5																																																									
FSMA ii	10	C3	5	C3 (&C4) 15																																																										
		C4	5	C5	20																																																									
FSMA v	10	C3	5	C6	5																																																									
		C6	5	C7	10																																																									
FSMA iii	10	C5	10																																																											
FSMA iv	10	C5	10																																																											
FSMA vi	10	C7	10																																																											
Total	60		60																																																											
Reviewer #2	See my comment in 3.																																																													
Reviewer #3	Related to this subject, decision analysis centers on the idea that by quantifying the preferences for each <i>criteria</i> rather than specific food, a more objective and systematic prioritization framework can be achieved. By defining which criteria are most important, and integrating these respective weights with scores representative of foods performance by each criteria, an integrated risk score can be quantified.																																																													

CHARGE QUESTION 4: <i>Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.</i>		
REVIEWER	COMMENT	RESPONSE
	There are different ways to elicit weights and to assign scores, some of which are tailored to specific MCDA methodologies. In general, if the individual hazard scores represent order-of-magnitude risk factors, equal weighting is most appropriate from a mathematical perspective. If the individual scores represent additive indicators of some sort, then different weights could be used, where they would provide flexible scaling factors.	
Reviewer #4	<p>Before commencing the response to this specific question, it is worth noting that Figure 6-1 in the Draft Report presents a distribution for all product:hazard pairs. In the Draft Report it is suggested that useful information could be derived from the proportion of cases above, or below, certain “threshold” scores, e.g. that appear as inflection points on the plot. To place that notion in context, I developed a simple stochastic model using Analytica stochastic simulation software. The model had seven variables, each of which could have a value of 0, 1, 3 or 9. Equal probability (0.25) was assigned to each value for each of the seven variables. The model calculated the sum of the seven values. 1286 iterations of the model were executed, and the score for each iteration recorded. The scores were sorted, and then graphed, with the same axes as Figure 6-1 of the Draft Report. That plot is presented overleaf (Figure A) and was developed to show that the shape of the plot is strongly influenced by the number of ways that intermediate scores can be achieved, compared to very high, or very low values. The similarity of the graph based on combinations of completely randomly derived values (Figure A), compared to the outputs of the model based on its ascribed values derived from data and expert opinion (Figure 6-1) suggests that caution needs to be exercised if using Figure 6-1 to derive “threshold” values for the designation of a HRF and, in particular, if intending to ascribe special meaning to inflection points on the plot as is hinted at in Section 6.2 of the RRM Draft report.</p>	

CHARGE QUESTION 4: *Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.*

REVIEWER	COMMENT	RESPONSE
	<div data-bbox="531 354 1507 927" data-label="Figure"> </div> <p data-bbox="457 971 1623 1084">Figure A. Results of a simple stochastic model emulating the risk scoring system and showing the distribution of food risk scores that would arise from a <i>completely random</i> allocation of scores for each of seven criteria, illustrating the strong similarity to the scores distribution derived from all food:hazard pairs considered in the RRM.</p> <p data-bbox="382 1166 1640 1414">At L.18 <i>ff.</i> on p. 85, it is noted that food:hazard pairs with a score of >310 have a majority of scores of 3 or 9 (in the absence of additional weighting, i.e., all weights equal to 10). This observation is a simple consequence of the scoring scheme itself. The additive seven-criterion scoring scheme is based on discrete score values of 0, 1, 3 or 9 for each of the seven variables. From that it is impossible to generate a score of 310 unless at least 3 of the scores are 9 and one of the scores is 3, i.e., the majority of the seven criterion scores. Thus, the discussion in this section of the Draft Report seems irrelevant and inane. The observations discussed in that section</p>	

CHARGE QUESTION 4: *Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.*

REVIEWER	COMMENT	RESPONSE
	<p>are about the consequences of the scoring scheme, not the underlying processes and phenomena that dictate the food safety risk, and which the scoring scheme/risk ranking model should aim to reflect.</p> <p><i>Criteria weighting</i></p> <p>Having conducted and been involved in several fully quantitative risk assessments, I can attest that I have no intuitive sense of which risk-affecting factors would be most important in a risk-ranking process, and should therefore be ascribed additional weight. Importantly, the factors included explicitly in a fully quantitative risk assessment do not readily correlate with the ‘compound’ variables included in the Draft Risk Ranking model under review. Similarly, the influence of a factor can vary according to the extent of variability defined for that factor and, if the variability in a factor changed significantly (e.g., the storage time for a product), that factor’s influence on the risk estimate might also change and require more (or less) weight to be assigned to it in a semi-quantitative risk assessment scheme such as presented in the Draft Report. The issue only arises, however, if a scoring scheme, rather than a fully quantitative risk assessment model is used for generating the risk estimates.</p> <p>I’d expect that different stakeholders would place more weight on certain kinds of information that they were more familiar with and that provided useful insights to them and their risk assessment, e.g., a food processor might weight product formulation/packaging and the rigor of the manufacturer’s HACCP system more heavily than epidemiological data, because outbreaks only occur when there are system failures. Conversely, an epidemiologist might rely on disease statistics more heavily because they have less understanding of processing and distribution factors that might have affected risk, instead relying more on the types and quantities of food eaten and the relative susceptibility of the victims. Neither approach is inherently better or worse. Any approach based on stakeholder opinion/experience would need to select a wide range of expert stakeholders and also seek to develop a consensus approach. The consensus process might</p>	

CHARGE QUESTION 4: *Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.*

REVIEWER	COMMENT	RESPONSE
	<p>also need to consider the reliability of the various data sources used to provide estimates for the factors/criteria in the Draft Model.</p> <p>As the discussion in Appendix M1 of the draft RRM report illustrated, there is no simple answer to the question of weighting, nor any single approach that is widely endorsed, despite that several apparently rigorous approaches have been articulated.</p> <p>As stated earlier, as a risk modeler my preferred solution would be to use a model that explicitly considers all the fundamental factors that affect food safety risks, and to unambiguously describe the interrelationships between those factors, to develop rigorous and transparent risk estimates and rankings. Doing so would obviate the need for weighting factors.</p> <p>Failing that, another approach implicitly suggested in Appendix M1 would be to conduct sensitivity analyses on the model to better understand which criteria had most influence. A problem with that approach is that the influence that those criteria already have may be inappropriate.</p> <p>Another approach would be to conduct sensitivity analyses on existing fully quantitative risk assessment models to see which factors, or combinations of factors relevant to criteria used in the Draft Risk Ranking model have the most influence in the risk estimates generated and to derive appropriate weighting factors from that analysis.</p> <p>Having experimented with the RRM model (i.e., via the Access model interface) and thought about relative weightings for Criteria and specifically the reliability of epidemiological information and illness severity information - as opposed to the inferences about risk generated by the RRM - I began to wonder whether the “epidemiological link” Criterion and “severity” Criterion should be weighted MUCH more heavily than Criteria 3 to 6 which aim to infer the risk (essentially the same information as the epidemiological data), rather than relying solely on the (imperfect) epidemiological data. Thus criterion 1 might receive 30% weight, Criterion 2 would</p>	

CHARGE QUESTION 4: *Considering the five different criteria weighting schemes described in Sections 6.2 and 6.3, are any one of these schemes (equal and non-equal weighting) not appropriate to consider for the intended purpose? For example, what weighting scheme is most useful? What weighting scheme should be avoided? Please make any additional recommendations on weighting schemes that might be considered for the proposed criteria and criteria indicators. Please explain the rationale behind your suggestions.*

REVIEWER	COMMENT	RESPONSE
	<p>receive 20%, and Criteria 3 to 6 would receive 7.5% of the weight each, and Criterion 7 would receive 20%. In this manner the actual epidemiological data indicating frequency of illness would have the same weight as the inferred level of illness (inferred from Criteria 3 – 6) and both would be combined equally with the severity criteria (i.e., Criteria 2 and 7). Criteria 2 and 7 are given equal weight because they are in, in fact, different dimensions of same element of risk, (i.e., ‘severity’).</p> <p>The Codex Alimentarius Commission and other organizations that have proffered guidelines for food safety risk assessment have identified a number of fundamental criteria for the reliable conduct of such risk assessment. Common to most of these guidelines/recommendations is that the process should be “science-based”, and that the process used, including assumptions and limitations, should be “transparent”, i.e., that all data sources, assumptions, uncertainties and limitations and their consequences should be clearly and thoroughly documented as part of the risk assessments. These desiderata arise because perfect data to support risk assessment are seldom available, i.e., risk assessments usually represent only the best estimates available within the constraints and limitations of the data and current knowledge and those limitations, and their potential consequences for risk management decisions and the stakeholders affected by them, need to be made clear.</p>	
Reviewer #5	[To be completed]	

CHARGE QUESTION 5: *Considering the various scenarios (described in Section 7.3 and Appendix N) to aggregate food-hazard pairs in order to identify the foods which should be identified as high-risk vs. not high-risk, which option(s) are more appropriate to consider and why?*

REVIEWER	COMMENT	RESPONSE
Reviewer #1	<p>What constitute/define “high-risk food” would ultimately have an impact on what aggregation approach is used to derive composite scores for foods for rank ordering. High risk could be based on: 1) frequency of hazards in a single food (i.e., no. of hazards in a given food), or 2) when hazard occurs in a food it is a very high risk (i.e., high FRS score), or 3) a combination of both (1) and (2).</p>	

CHARGE QUESTION 5: Considering the various scenarios (described in Section 7.3 and Appendix N) to aggregate food-hazard pairs in order to identify the foods which should be identified as high-risk vs. not high-risk, which option(s) are more appropriate to consider and why?

REVIEWER	COMMENT	RESPONSE
	<p>In the various analyses provided in the draft report Table 7-4 as well as in appendix N, FDA attempted to tackle these definitions by various aggregation schemes, i.e., summation, average, maximum FRS, or cutoff based on FRS score and sum, etc. Overall, the current approach as described in Section 7, which is a cutoff for FRS at 270 prior to summing to derive composite score for foods and then rank, is defining “high-risk” food based on (2), i.e., when hazard occurs in a food it is a very high risk (high FRS score above a set cutoff). This approach deliberately dismisses foods with multiple hazards but not of high FRS. While it is not unreasonable to not allow lower risk hazard to have high influence, it is possible that foods with multiple small problems could be considered “high risk” food and this approach could be viewed negatively by some.</p>	
Reviewer #2	<p>It is hard to understand all the rationale behind each tested approach. Alternative 1: not acceptable, for the same reason I explained previously for the aggregation of the seven criteria. Alternative 2: the compensation is not acceptable. High score with one hazard will be compensated by low score with another hazard. Can we say that one food is on average safe or on average at high risk? Alternative 3: it is one possible option to reduce the number of hazards per food. Alternative 4: this option will underestimate the role of food with several hazards. Not acceptable Alternative 5: Not acceptable. Because we are mixing factors that have not the same impact in regard to the different hazards.</p>	
Reviewer #3	See response under 5.a.	
Reviewer #4	<p>The risk rankings presented in Appendix N, based on various methods of aggregation, include results that are ‘unexpected’. The discussion included in Section N and the draft report also implicitly suggest that stakeholders, and perhaps even the model developers, consider that the rankings may not accurately reflect the true risks, e.g., on p. 2 of Appendix N it is stated “concerns were raised that perhaps there were potential biases in the model”.</p> <p><i>While it is entirely possible for a risk assessment process to produce unexpected results, and for those unexpected results to be challenged by stakeholders and experts who hold different perceptions of relative risks, those model predictions may nonetheless be valid.</i> The resolution of these apparent discrepancies should involve exploration of the logic of the model to determine whether the unexpected results can be understood (and represent scenarios and phenomena that may not yet have been perceived) or whether there are simply logical errors in the model.</p>	

CHARGE QUESTION 5: Considering the various scenarios (described in Section 7.3 and Appendix N) to aggregate food-hazard pairs in order to identify the foods which should be identified as high-risk vs. not high-risk, which option(s) are more appropriate to consider and why?

REVIEWER	COMMENT	RESPONSE
	<p>Also, while the available epidemiological data are less than perfect, they do provide a point of reference by which to evaluate the credibility of the risk-ranking model - at least for product:hazard pairs that are part of the reporting/surveillance system. If the model rankings do not reflect the <i>known</i> situations for which reliable data are available, it suggests that there may be errors in the model.</p> <p>An obvious problem with rankings based on the <i>sum</i> of the scores for each hazard considered to be associated with a food is that the attribution of more hazards, even if trivial, still adds to the score because some hazards have high scores for some criteria even if their likelihood of occurrence is negligible. Similarly, as noted earlier, the logic of the model doesn’t have a way of effectively eliminating the risk from trivial hazards, e.g., due to cooking before eating, lethal treatments applied in hermetic packages, etc. As commented in the complementary review of the data upon which the RRM is based, using the total sum approach requires that every hazard needs to be assessed for every product. A specific example illustrating this is presented below).</p> <p>Similarly, in Appendix N an averaging process was shown to distort the rankings because a food that has been associated with a hazard that is trivial, and another that has strong evidence as a persistent source of human illness will be under-estimated, even though it is logically clear that a real risk cannot be reduced by a lesser risk. The risk to human health from a single product is the sum of the risks from each associated hazard, even if some of the hazards are trivial.</p> <p>The preferred option in Appendix N was to calculate the sum of the highest score for each criterion from any hazard considered to be associated with the product. <i>This is a completely illogical approach and is not scientifically defensible</i> because it involves mixing of attributes of one hazard with those of another, i.e., <i>the approach completely undermines the logic that the model is based on</i> and that such an approach should even be considered as a means of reconciling the model predictions with stakeholder expertise and perceptions suggests problems with the logic/structure of the model.</p> <p>At least some of these problems are possibly due to incorrect approaches to combining variables that scale exponentially, and those that scale arithmetically, as more fully discussed in response to Q5a, below.</p> <p>As an aside, at L. 20-22 on p. 107 of the draft report, it is stated that: “the risk ranking model might not capture the higher risk food-hazard pairs involving chemical hazards and undeclared allergens, because these agents do not grow in food and Criterion 4 is by default 0”. The basis of this comment requires some</p>	

CHARGE QUESTION 5: *Considering the various scenarios (described in Section 7.3 and Appendix N) to aggregate food-hazard pairs in order to identify the foods which should be identified as high-risk vs. not high-risk, which option(s) are more appropriate to consider and why?*

REVIEWER	COMMENT	RESPONSE
	<p>explanation, and link to the evidence base that suggests that the risk from these hazard groups is underestimated. However, examination of the risk ranking tables suggests that risk of allergens is frequently ranked very highly (see also my comments in the complementary data review), again suggesting that the basis of the comment needs further explanation/justification/investigation as to its validity. This issue is discussed further under Question 6b by reference to independent estimates of the relative burden of foodborne disease from different classes of hazards presented by WHO in late 2015.</p>	
Reviewer #5	[To be completed]	

CHARGE QUESTION 5a: *Are there additional aggregation method(s) that might be considered? Please explain.*

REVIEWER	COMMENT	RESPONSE
Reviewer #1	<p>Some thoughts should be given further into the cutoff point for the FRS. Various cutoff point should be considered in a sensitivity analysis. If microbial, chemical, and allergens are treated equally in the raw scores for each criterion, then there is no need for a different cutoff point. However, if the current unbalance score is kept (i.e., microbes have an additional C4 score of up to 9), then a different cutoff (lower) would have to be considered for chemical and allergens.</p>	
Reviewer #2	<p>Another option is to use ELECTRE III to rank the food. The criteria will be the ranks obtained for each hazard. The pairwise comparison will allow calculating concordance and discordance indices based on the different ranks obtained for each hazard. The discrimination thresholds will be the distance between ranks. Then we can define strict difference if for example the difference between ranks is higher than 10. When a hazard is not considered for one food, the advantage will be given to the other food. However to avoid counting discordances corresponding to hazard ranked down a threshold may be chosen (i.e., < 100 to be included as advantage if the food is not concerned by a given hazard).</p>	

CHARGE QUESTION 5a: Are there additional aggregation method(s) that might be considered? Please explain.																																																																																				
REVIEWER	COMMENT	RESPONSE																																																																																		
	<p>Example of pairwise comparison: Just an example without taking into account for discrimination thresholds.</p> <table border="1"> <thead> <tr> <th></th> <th colspan="2">Dairy - Ice Cream and Related</th> <th colspan="2">Seafood - N.E.C.</th> </tr> <tr> <th>Hazard</th> <th>Advantage</th> <th>Rank</th> <th>Advantage</th> <th>Rank</th> </tr> </thead> <tbody> <tr> <td><i>Campylobacter</i> spp.</td> <td>0</td> <td></td> <td>1</td> <td>37</td> </tr> <tr> <td>Ciguatoxin</td> <td>0</td> <td></td> <td>1</td> <td>55</td> </tr> <tr> <td><i>Clostridium botulinum</i></td> <td>0</td> <td></td> <td>1</td> <td>75</td> </tr> <tr> <td>Hepatitis A virus</td> <td>1</td> <td>19</td> <td>0</td> <td>20</td> </tr> <tr> <td><i>Listeria monocytogenes</i></td> <td>1</td> <td>13</td> <td>0</td> <td>95</td> </tr> <tr> <td>Norovirus</td> <td>1</td> <td>28</td> <td>0</td> <td>69</td> </tr> <tr> <td><i>Salmonella</i> spp.</td> <td>1</td> <td>15</td> <td>0</td> <td>44</td> </tr> <tr> <td>Scombroid toxin (Histamine)</td> <td>0</td> <td></td> <td>1</td> <td>55</td> </tr> <tr> <td><i>Shigella</i> spp.</td> <td>0</td> <td></td> <td>1</td> <td>32</td> </tr> <tr> <td><i>Staphylococcus aureus</i></td> <td>1</td> <td>25</td> <td>0</td> <td>35</td> </tr> <tr> <td>STEC O157</td> <td>1</td> <td>14</td> <td>0</td> <td>114</td> </tr> <tr> <td>Undeclared allergens</td> <td>0</td> <td></td> <td>1</td> <td>55</td> </tr> <tr> <td>Undeclared allergens (other than fish)</td> <td>1</td> <td>5</td> <td>0</td> <td>20</td> </tr> <tr> <td><i>Vibrio parahaemolyticus</i></td> <td>0</td> <td></td> <td>1</td> <td></td> </tr> </tbody> </table> <p>If you are interested by the idea, I can provide more details.</p>		Dairy - Ice Cream and Related		Seafood - N.E.C.		Hazard	Advantage	Rank	Advantage	Rank	<i>Campylobacter</i> spp.	0		1	37	Ciguatoxin	0		1	55	<i>Clostridium botulinum</i>	0		1	75	Hepatitis A virus	1	19	0	20	<i>Listeria monocytogenes</i>	1	13	0	95	Norovirus	1	28	0	69	<i>Salmonella</i> spp.	1	15	0	44	Scombroid toxin (Histamine)	0		1	55	<i>Shigella</i> spp.	0		1	32	<i>Staphylococcus aureus</i>	1	25	0	35	STEC O157	1	14	0	114	Undeclared allergens	0		1	55	Undeclared allergens (other than fish)	1	5	0	20	<i>Vibrio parahaemolyticus</i>	0		1				
	Dairy - Ice Cream and Related		Seafood - N.E.C.																																																																																	
Hazard	Advantage	Rank	Advantage	Rank																																																																																
<i>Campylobacter</i> spp.	0		1	37																																																																																
Ciguatoxin	0		1	55																																																																																
<i>Clostridium botulinum</i>	0		1	75																																																																																
Hepatitis A virus	1	19	0	20																																																																																
<i>Listeria monocytogenes</i>	1	13	0	95																																																																																
Norovirus	1	28	0	69																																																																																
<i>Salmonella</i> spp.	1	15	0	44																																																																																
Scombroid toxin (Histamine)	0		1	55																																																																																
<i>Shigella</i> spp.	0		1	32																																																																																
<i>Staphylococcus aureus</i>	1	25	0	35																																																																																
STEC O157	1	14	0	114																																																																																
Undeclared allergens	0		1	55																																																																																
Undeclared allergens (other than fish)	1	5	0	20																																																																																
<i>Vibrio parahaemolyticus</i>	0		1																																																																																	
Reviewer #3	<p>Depending on what the food-hazard pair scores represent, summing them to identify high-risk foods may or may not be an appropriate step. If they represent order-of-magnitude risks, then summing food-hazard pair scores would be inappropriate and a more detailed aggregation rule would be necessary. If they represent magnitudes of risks, then summing them, with some weighting corresponding to relative quantities consumed of the different foods, would be the appropriate course of action.</p>																																																																																			
Reviewer #4	<p>Yes. As discussed earlier in this report, the risk ranking scheme involves combination of variables that scale exponentially, and the risk scoring process essentially produces a ranking on an exponential scale. However, if a product has risks attributable to multiple hazards, the overall risk is the sum of the risks, not the product of the individual risks. To explain, adding risk scores that are based on exponential scales is</p>																																																																																			

CHARGE QUESTION 5a: Are there additional aggregation method(s) that might be considered? Please explain.		
REVIEWER	COMMENT	RESPONSE
	<p>logically equivalent to multiplying the risks from each hazard, rather than adding them together to achieve an overall risk estimate. Given this, a more appropriate approach to assessing the combined risk from multiple hazards is to convert the risk estimates to an arithmetic scale, add the risks together, and then convert the sum back to an exponential scale, i.e.,</p> <p>Aggregate Risk Score, $RS_{total}F$ for food, F, with risk scores for each hazard given by $RS_1F, RS_2F \dots RS_nF$ for n identified hazards is:</p> $RS_{total}F = \log_{10}(10^{RS_1F} + 10^{RS_2F} + 10^{RS_3F} + \dots + 10^{RS_nF})$ <p>As an example of why this approach is necessary, and one that will probably be familiar to people working in microbial food safety, consider a food that is initially contaminated with microorganisms, but is then further contaminated by contact with a contaminated surface. If the concentration of the organisms on the food is 10^2CFU/g and the additional contamination is 10,000 (10^4) organisms, and the food weighs 100g, the final concentration is $(10^2 \text{ CFU/g} * 100\text{g} + 10^4 \text{ CFU})/100\text{g} = 20,000/100\text{g} = \log_{10}2.3\text{CFU/g}$. i.e., despite that the additional cross contamination is the equivalent of 10^2CFU/g, the final contamination is <i>not</i> 10^{2+2}CFU/g but 10^2+10^2CFU/g.</p> <p>Using this (more) logically defensible approach, the cumulative risk score would scale more naturally, and will be less affected by differences in the total number of hazards ascribed to the various foods.</p> <p><i>An overt logical error was discovered in the model, however, relating to aggregation irrespective of the treatment or weighting applied that further compounds the anomalies.</i></p> <p>To illustrate the problem, in the assessment of the microbial risks associated with finfish, 16 microbial hazards were identified and evaluated in the risk ranking, leading to a relative risk of ~400. In a separate category ‘finfish associated with histamine’ was associated with only one hazard, i.e., histamine, and generated a risk score of only 25. This is illogical. <i>Every other microbial hazard associated with finfish can also be associated with finfish capable of developing high histamine levels.</i> Epidemiological evidence shows that histamine intoxication from scombroid fish, while usually not severe, is the commonest cause of illness related to finfish. As such, the risk from histamine producing finfish is systematically underestimated because of this (arbitrary?) association of hazards with different types of finfish, as was suggested above.</p>	

CHARGE QUESTION 5a: Are there additional aggregation method(s) that might be considered? Please explain.		
REVIEWER	COMMENT	RESPONSE
	<p>More importantly, however, this treatment of the hazards data reveals a systematic logical error in the model that was alluded to (in the discussion above) about the cumulative effect of multiple hazards. In the scoring, every individual hazard score includes a score for ‘consumption’. Therefore, every additional hazard deemed to be associated with a product adds to the ‘consumption’ score again, even though the presence of an additional hazard does not change the consumption, it simply means more hazards are associated with the <i>same amount</i> of consumption. In the logic of the scoring scheme, however, each additional hazard effectively assumes that the consumption is increased. In the case of finfish, with a consumption score of ‘3’, the sixteen hazards identified as being associated with finfish mean that the contribution of consumption is scored as 48 instead of 3, on a scale where the maximum consumption score should be ‘9’. Presumably, this error is inherent throughout the relative risk scores where more than one hazard is associated with a food or food category. As such, the more hazards deemed to be associated with the hazard, the more inflated the consumption score becomes and the more distorted the relative ranking becomes, confirming comments above about a correct method of aggregation of scores from multiple hazards. Currently, the method is logically incorrect. Similarly, where the ‘at risk’ population is only a subset of the population, the ‘consumption’ values have to be modified accordingly (as was discussed in the complementary report on the model data).</p>	
Reviewer #5	[To be completed]	

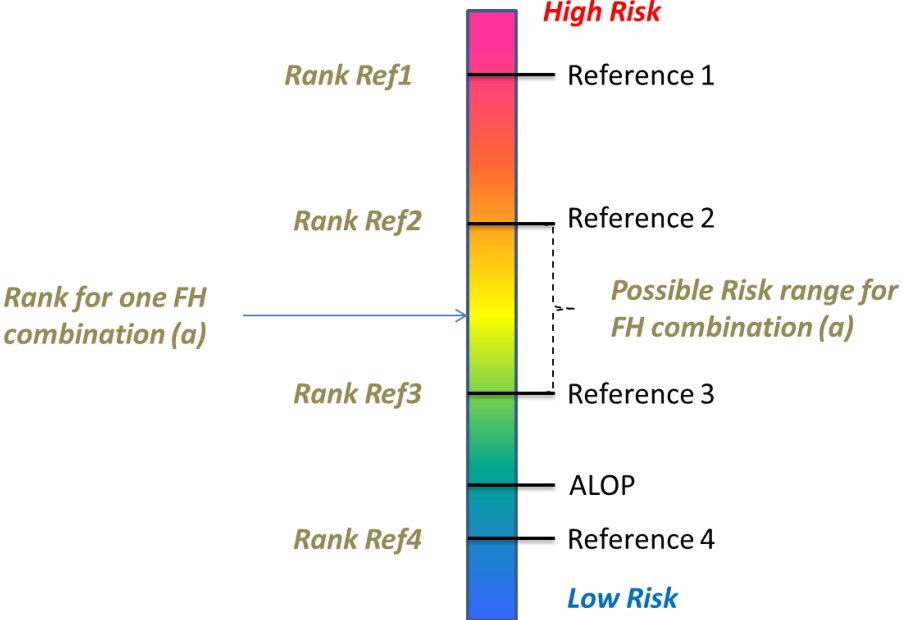
CHARGE QUESTION 6: Given the underlying data supporting the scoring, what are the considerations to take into account when identifying high risk vs. not high risk food-hazard pairs or foods?		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	<p>Based on the appendix J, the percent consumer is based on the entire population. Considerations should be made to incorporate this into criterion 6 to consider sensitive subpopulations such as children, pregnant women, and the elderly.</p> <p>Also for the economic criterion 7, some discussions, considerations with respects to economic burden to society when the effect is on child/fetus should be provided (it is unclear if these considerations were accounted for in the scoring of this metric in the current model).</p> <p>The uncertainty/confidence in the scores are tracked. However, it is unclear as to how this information will be used in ranking of high risk foods. There need to be some discussions of what the Agency intends to do with areas where there is lack of confidence of the data and results are highly uncertain.</p>	
Reviewer #2	The system needs to be calibrated using external data. See my next comment.	

CHARGE QUESTION 6: Given the underlying data supporting the scoring, what are the considerations to take into account when identifying high risk vs. not high risk food-hazard pairs or foods?		
REVIEWER	COMMENT	RESPONSE
Reviewer #3	Perhaps there could be two different thresholds in use, including (i) one separated at the level of hazard type, and (ii) one collectively for the food. For example, either a collective score above 300, or a score on any of the three separate hazards above 200 might qualify as a high risk that merits special handling. The rationale for the former is what seems to underlie the whole approach, simply as an indicator of the overall danger. The rationale for the latter might be that emergent conditions related to the offending hazard could inflate the hazard to more worrisome levels, i.e., the food's risk estimate might have “fatter tails” if it is driven by one of the hazards.	
Reviewer #4	<p>There are numerous factors that influence food-borne risk, as discussed above in response to Question 1. These include:</p> <ul style="list-style-type: none"> • the potential for contamination with a hazard at a level likely to cause human illness, or the potential during normal handling to increase (grow) to a level that could cause illness (considering also the proportion of the population susceptible to the hazard), • the normal use of the product (e.g., RTE or cooked before consumption), • the existence of reliable CCPs for those hazards and their reliable implementation, or quality assurance systems that reliably detect contamination and thereby effect removal of contaminated lots from distribution, • whether the product can be recontaminated with a disease-causing dose <i>after</i> a reliable CCP treatment/action or quality assurance process, • the severity of the symptoms associated with disease caused by consumption of the hazard, • the frequency of consumption of the food (whether by individuals, or total consumption by population, or only by the susceptible population, as relevant), • whether the hazard accumulates in the body of the consumer or whether each exposure is a discrete event (and whether earlier exposures provide protection against subsequent exposures, i.e., immunity, or predispose the consumer to more severe symptoms upon subsequent exposure, e.g., induced hypersensitivity), etc. <p>As noted earlier, all these elements of risk are implicit (though perhaps confounded) in the six factors nominated for inclusion by FDA, and are translated in various combinations into the seven criteria adopted for the draft Risk Ranking Model (RRM). Consumption was not explicitly considered in the six factors identified by FDA, however, but that omission may be appropriate if the risk is intended to be ranked on a per-serving basis. The likelihood of a disease-causing dose also is not considered explicitly (nor implicitly as far as I can determine) and this seems to be a weakness in the draft RRM.</p>	

CHARGE QUESTION 6: <i>Given the underlying data supporting the scoring, what are the considerations to take into account when identifying high risk vs. not high risk food-hazard pairs or foods?</i>		
REVIEWER	COMMENT	RESPONSE
	Interestingly, the FDA's decision criteria/logic for identification of Potentially Hazardous Foods/Temperature Controlled for Safety Foods seems to be relevant to identification of "riskier" foods (i.e., at least those subject to potential microbial contamination). Thus, that evaluation system might have provided a useful and consistent starting point for the RRM which is intended to identify "high-risk foods that require additional record keeping".	
Reviewer #5	[To be completed]	

CHARGE QUESTION 6a: <i>What are the pros and cons in establishing a threshold considering all three types of hazards in the model, vs. drawing a line separately for microbial, chemical and undeclared allergens in foods?</i>		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	See comment above in question 5.	
Reviewer #2	A threshold without external data is actually difficult. One option is to consider a set of food-hazard combinations for which quantitative data and quantitative risk assessment are available and use them to create a sort of references to be compared with the U.S. ALOP (acceptable level of protection, i.e., 1/million DALYS per year per consumer)... it will be a way to have a sort of risk graduations:	

CHARGE QUESTION 6a: What are the pros and cons in establishing a threshold considering all three types of hazards in the model, vs. drawing a line separately for microbial, chemical and undeclared allergens in foods?

REVIEWER	COMMENT	RESPONSE
		
<p>Reviewer #3</p>	<p>Perhaps there could be two different thresholds in use, including (i) one separated at the level of hazard type, and (ii) one collectively for the food. For example, either a collective score above 300, or a score on any of the three separate hazards above 200 might qualify as a high risk that merits special handling. The rationale for the former is what seems to underlie the whole approach, simply as an indicator of the overall danger. The rationale for the latter might be that emergent conditions related to the offending hazard could inflate the hazard to more worrisome levels, i.e., the food's risk estimate might have “fatter tails” if it is driven by one of the hazards.</p>	
<p>Reviewer #4</p>	<p>Ideally, there would be a simple, transparent and objective risk evaluation system that accurately estimates, and ranks, the overall risk from a diversity of microbial, chemical and allergen hazards that could contaminate the identified foods of interest. However, the variety of food composition, processing and handling steps, the responses of the hazards to those treatments, including over time, the type of effects of the hazards on consumers, and the strong differences in consumer susceptibility, seem to</p>	

CHARGE QUESTION 6a: <i>What are the pros and cons in establishing a threshold considering all three types of hazards in the model, vs. drawing a line separately for microbial, chemical and undeclared allergens in foods?</i>		
REVIEWER	COMMENT	RESPONSE
	<p>preclude such a system being feasible, unless supported by much more data than is currently available to assess each of the relevant factors.</p> <p>As suggested in earlier comments there appear to be anomalies in the preliminary risk rankings presented, with estimated risk from undeclared allergens appearing to be dominant despite the lack of confirming, or even supporting, epidemiological evidence. However, ‘absence of evidence is not evidence of absence’ and there may be a high burden of (unreported) illness from allergens in the U.S. food supply. The same is true of chemical hazards in the U.S. food supply, i.e., there is a paucity of relevant data.</p> <p>However, while it was not available at the time of preparation of the draft RRM, the World Health Organization report on the Global Burden of Foodborne Disease (released in December 2015; http://www.who.int/foodsafety/areas_work/foodborne-diseases/ferg/en/) provides an independent, expert and well-resourced evaluation of relative risk from different types of hazards in foods. The WHO report considers the burden of foodborne illness in different regions of the world including the ‘AMR’ region comprising Canada, U.S.A. and Cuba. The population of Cuba is ~11 million. The population of Canada is ~36 million. The population of U.S.A. is 323 million. As such, the estimated burden of disease in AMR will be dominated by U.S. statistics.</p> <p>Collectively, Tables A8.4, A8.6, and A8.7 in the WHO report permit estimates of the disease burden from selected chemicals, peanut allergens, and a plethora of microbial (bacterial, protozoal and fungal pathogens and their toxins) hazards. Due to data limitations, however, the only ‘chemical’ hazards considered were aflatoxin, dioxin and cyanide from Cassava. Clearly, there is negligible risk from Casava consumption in North America.</p> <p>Table A.6 in the WHO report considers the burden of foodborne illness in: i) the north America Region, ii) European region and iii) Western Pacific Region (a mix of 37 countries including affluent ‘westernized’ nations and developing island nations). In those regions the combined burden from peanut allergens was ~10% of the burden from dioxin and aflatoxin combined. Given that the Western Pacific region includes many tropical nations, the relative risk from aflatoxin in those countries may be relatively higher, thereby reducing the apparent relative importance of peanut allergens.</p> <p>Nonetheless, the estimated DALY burden per 100,000 population in those three regions for aflatoxin and dioxin combined was 2 (95% CI: 0.6-24), while the DALY burden for all microbial hazards was 51 (95% CI: 41–112). Based on the statistic above, the relative DALY burden per 100,000 for peanut allergens</p>	

CHARGE QUESTION 6a: What are the pros and cons in establishing a threshold considering all three types of hazards in the model, vs. drawing a line separately for microbial, chemical and undeclared allergens in foods?		
REVIEWER	COMMENT	RESPONSE
	<p>would be ~0.2. (Peanut allergies are considered to be the most common in the ‘developed’ world). These relative burden of disease estimates do not seem to accord with the relative risks identified in the draft model for review, suggesting that relative risks attributed to microbial hazards, chemical hazards and allergens in the draft model are not consistent. The discrepancy between risk estimates for allergens in the draft RRM and Minor <i>et al.</i> (2015, Risk Anal., 35:1135-1139) was noted earlier. It is noted that chemical hazards and allergens have separate scoring schemes, especially for severity, and that relative susceptibility (or proportion of the population at risk) is not explicitly considered in the model.</p> <p>Given these considerations and the apparent discord between the WHO estimates and those from the draft RRM, it is this reviewer’s opinion that separate thresholds are needed if using the current RRM for determination of foods that require additional record-keeping due to possible high risk of: i) microbial hazards, ii) chemical hazards and iii) allergens. Further data are needed to reliably determine relative risks from these three categories of hazard. Alternatively, revision of the criteria scores for each type of hazard may be required to generate risk estimates that are comparable on a burden-of-disease basis.</p>	
Reviewer #5	[To be completed]	

CHARGE QUESTION 7: Are the seven criteria and scoring definitions implemented appropriately in the SAS codes (Appendix P) and the Access Model?																																																		
REVIEWER	COMMENT	RESPONSE																																																
Reviewer #1	It should be noted that I am not a SAS programmer and my SAS experience is limited and dated. With this in mind, I reviewed the appendix P as requested and provided some comments below.																																																	
Reviewer #2	<p>The codes are correct. However, there are some differences in the score in the Excel files and the final score in Access tables.</p> <p>Just one example of differences:</p> <table border="1"> <tbody> <tr> <td>19</td> <td>Fluid white milk, Grade A, pasteurized</td> <td>88</td> <td><i>Campylobacter</i> spp.</td> <td>5</td> <td>1</td> <td>9</td> <td>1</td> <td>4</td> <td>2.</td> <td>201</td> <td>186</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>8</td> <td>1</td> <td>5</td> </tr> <tr> <td>35</td> <td>Fresh herbs - Group</td> <td>231</td> <td><i>Cyclospora cayetanensis</i></td> <td>1</td> <td>3</td> <td>9</td> <td>1</td> <td>5</td> <td>3.</td> <td>201</td> <td>159</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td>2</td> <td></td> <td></td> <td></td> <td></td> <td>8</td> <td>3</td> <td>3</td> </tr> </tbody> </table>	19	Fluid white milk, Grade A, pasteurized	88	<i>Campylobacter</i> spp.	5	1	9	1	4	2.	201	186										8	1	5	35	Fresh herbs - Group	231	<i>Cyclospora cayetanensis</i>	1	3	9	1	5	3.	201	159					2					8	3	3	
19	Fluid white milk, Grade A, pasteurized	88	<i>Campylobacter</i> spp.	5	1	9	1	4	2.	201	186																																							
									8	1	5																																							
35	Fresh herbs - Group	231	<i>Cyclospora cayetanensis</i>	1	3	9	1	5	3.	201	159																																							
				2					8	3	3																																							
Reviewer #3	This reviewer cannot provide details on SAS codes and Access model since he has experience in risk assessment and decision analysis and does not have advanced knowledge in SAS and Access.																																																	
Reviewer #4	I do not have sufficient expertise in SAS and R, nor do I have sufficient expertise in M/S Access, to be able to make informed comment about the code underlying the RRM interface.																																																	

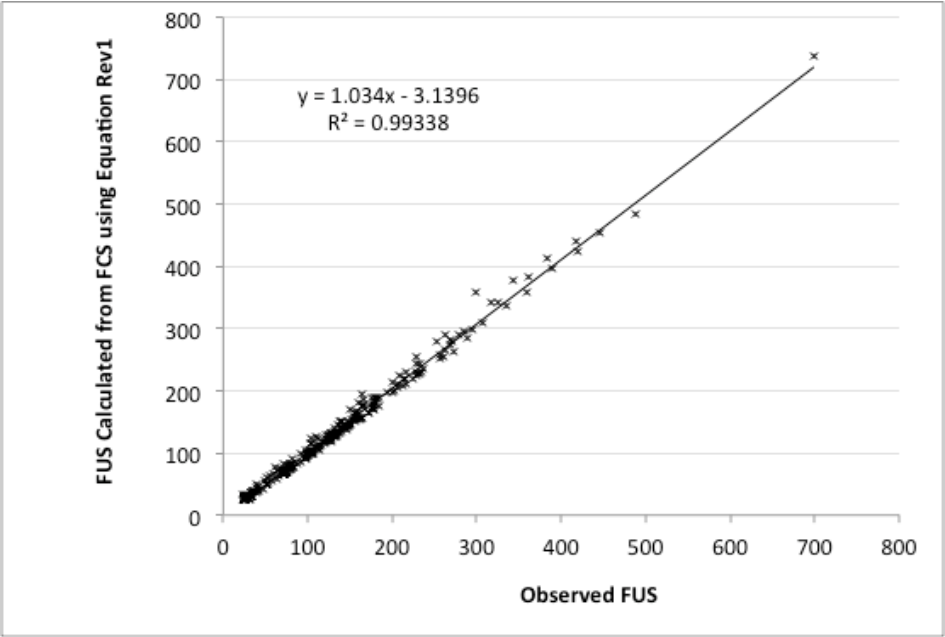
CHARGE QUESTION 7: Are the seven criteria and scoring definitions implemented appropriately in the SAS codes (Appendix P) and the Access Model?		
REVIEWER	COMMENT	RESPONSE
Reviewer #5	[To be completed]	

CHARGE QUESTION 7a.: Does the scoring logic described in Section 4 Figures 4-1 to 4-7) appropriately represent the scoring definitions described for each of the criteria in Section 3? If not, please describe what changes need to be made to correct it.		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	Yes, the scoring logic in Section 4 (Figures 4-1 to 4-7) appropriately represents the scoring definitions. Some questions/issues are provided above under question 2.	
Reviewer #2	Yes.	
Reviewer #3	This reviewer cannot provide details on SAS codes and Access model since he has experience in risk assessment and decision analysis and does not have advanced knowledge in SAS and Access.	
Reviewer #4	<p>Note. The criteria definitions were described in Section 2. Section 3 was concerned with identification of product:hazard pairs. Section 4 described the criteria definitions again, together with their implementation.</p> <p>Criterion 1, Section 2.2.1 text agrees with the scheme described in Figure 4-1.</p> <p>Criterion 2, Section 2.2.2 text agrees with the scheme described in Figure 4-2.</p> <p>Criterion 3, Section 2.2.3 text agrees with the scheme described in Figure 4-3.</p> <p>Criterion 4, Section 2.2.4. The text does not agree with the data/text in Figure 4-4 because no scores are given for the growth potential descriptors and shelf life descriptors at Section 2.2.4, while they are defined in Figure 4-4. Perhaps more importantly, the logic for combination of the two scores into a single score of 1, 3 or 9 is not described, i.e., there is a lack of transparency. Importantly, if the two scores are added, it appears that a value >10 is scored as 9, a value between 6 and 10 is scored as 3, and if the value is between 2 and 4 the value is scored as 1. If the values are multiplied, then a score (i.e., product) >27 and up to 81 is scored as '9'. Values of 9 are scored as 3 and values of 1 to 3 are scored as 1. By either approach, the relativities of the scale (and the risk contributions) are distorted by this process. The rationale for combination of the scores should be described clearly in the document.</p> <p>Criterion 5, Section 2.2.5. Same comments as above for Criterion 4 and Figure 4-4, i.e., no scores are given for the 'contamination probability during manufacturing' descriptors and 'steps taken to reduce contamination' descriptors at Section 2.2.4 but they are given in Figure 4-4. Perhaps more importantly, the</p>	

CHARGE QUESTION 7a.: Does the scoring logic described in Section 4 Figures 4-1 to 4-7) appropriately represent the scoring definitions described for each of the criteria in Section 3? If not, please describe what changes need to be made to correct it.		
REVIEWER	COMMENT	RESPONSE
	<p>logic for combination of the two scores into a single score of 1, 3 or 9 is not described, i.e., lack of transparency. The rationale for combination of the scores should be described clearly in the document.</p> <p>Criterion 6, Section 2.2.6 text does not agree with Figure 4-6 because Figure 4-6 includes scores for expert elicitation results that are not discussed at all in Section 2.2.6. Section 4.2.6. does mention that expert elicitation was used where specific data were not available from NHANES. Thus, the text at Section 2.2.6 is not fully consistent with Section 4.2.6.</p> <p>Criterion 7, Section 2.2.7 text does not agree with Figure 4-7 because Figure 4-7 includes scores from 'expert opinion' that are not discussed at all in Section 2.2.7. Section 4.2.7. does mention that expert elicitation was used where specific data were not available from other sources (i.e., Minor, . 2015; Scharf, 2011). Thus, the text at Section 2.2.7 is not fully consistent with Section 4.2.7.</p>	
Reviewer #5	[To be completed]	

CHARGE QUESTION 7b.: Are the scoring logic and order of preference accurately implemented in the SAS codes? (Please select 2-3 out of the 7 criteria for this evaluation). If not, please specify what changes need to be made.		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	<p>Criterion 1: appears to correctly implement the scoring of this criterion as described in Section 4, Figure 4-1</p> <p>Criterion 2: appears to correctly implement the upper portion of Figure 4-2 and Table 2-1. This reviewer cannot see where Table 2-2 and lower portion of Figure 4-2 (where expert judgment comes in for chemical and allergens). There are codes written for hazard id 73 (methanol), 2, and 3, but it is unclear why.</p> <p>Criterion 6: appears to be correctly implemented to tabulate percent consumers among U.S. population.</p>	
Reviewer #2	Yes.	
Reviewer #3	This reviewer cannot provide details on SAS codes and Access model since he has experience in risk assessment and decision analysis and does not have advanced knowledge in SAS and Access.	
Reviewer #4	I do not have sufficient expertise in M/S Access to be able to make informed comment about the accuracy of the SAS code underlying the RRM interface. However, I used the model in Access to generate ranking data for selected foods and hazard sets. By using the "Adjust Scores" dialog box I was able to identify the scores for each criterion for each hazard associated with the selected food and, from this, I was able to	

CHARGE QUESTION 7b.: Are the scoring logic and order of preference accurately implemented in the SAS codes? (Please select 2-3 out of the 7 criteria for this evaluation). If not, please specify what changes need to be made.		
REVIEWER	COMMENT	RESPONSE
	<p>check the simple additions (Eqn. 3). I then changed weighting combinations and re-ran the scenario and rechecked the additions leading to the FRS. I repeated this process for another set of weighting combinations. From this I was able to check the correctness of implementation of the weighting factors.</p> <p>The food:hazard pairs I considered were:</p> <ul style="list-style-type: none"> i) “Eggs” for all hazards and, from the four outputs, I selected “Egg dishes” (which encompassed 7 hazards) for further assessment by checking additions leading to the FRS, and additions after changing the weighting. ii) “Pasta -Dried Pasta” and “All Hazards”, which encompassed 8 associated hazards. iii) “Seafood - Finfish” for “Microbial Hazards” only and from the output selected “Finfish” but also “Finfish (Histamine Producing Species)” for more detailed assessment. For “Finfish” there were 16 discrete hazards, while for ‘Finfish (Histamine Producing Species) there was only one hazard, i.e., histamine. <p>In all cases, the additions and aggregated scores were correct and consistent with correct implementation of Equation 3. However, this assessment did reveal a logical error in the model that was described in response to Question 5a (regarding accumulation of consumption scores).</p> <p>Also, from data presented in the “Adjust Scores” dialog box the calculation of the uncertainty score was ‘checked’. The uncertainty score should be the simple sum of the individual uncertainty scores for each criterion for each hazard associated with the food. The definition of the uncertainty score in Section 4.4.2 and Equation 4 does not describe inclusion of weighting factors, but it was noted (through the ‘experiments’ described above) that the Uncertainty Score (and the Confidence Score) are also affected by the designation of the weighting factors specified in the ‘Ranking Criteria and Weights’ in the same way (algebraically) that the criterion scores are affected. Whether such weighting of the FCS and FUS is logically justified is debatable, but it certainly needs to be described in the documentation. Whether the omission of weighting factors in the definition of Eqn.4 is an oversight, or whether there is an error in the model code is not clear but a correction is required in one or the other.</p> <p>As an aside, there seems little value in including both the FCS and FUS as model outputs, as they effectively measure the same properties of the risk-ranking scores but on inverse scales, i.e., if the FSC for a criterion is 9, then the FUS must be 1, and <i>vice versa</i>, and when FCS is 3 so is the FUS. The net effect of</p>	

CHARGE QUESTION 7b.: Are the scoring logic and order of preference accurately implemented in the SAS codes? (Please select 2-3 out of the 7 criteria for this evaluation). If not, please specify what changes need to be made.		
REVIEWER	COMMENT	RESPONSE
	<p>this is that the two scores are strongly negatively correlated. To illustrate this, an approximation between these 'scores' is given by:</p> $\text{FCS} = 64.85 * \text{number of hazards associated with food} - \text{FUS (Eqn. Rev1)}$ <p>The observed FCS is plotted against the predicted FCS based on Eqn. Rev1 in Figure B, below.</p>  <p>Figure B. Results of a simple relationship (Eqn. Rev1) relating FCS to FUS, and showing the success of that relationship (regression and r^2 value shown) in predicting FUS from the FCS.</p> <p>On the basis of this analysis, the additional insights offered by the inclusion of both FSC and FUS in the model outputs should be explained.</p>	
Reviewer #5	[To be completed]	

CHARGE QUESTION 7c.: Are equations 1 through 4 and data weighting factors accurately implemented in the model (either the SAS codes or the Access Model)?		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	Equations 1 and 2 (for criterion 3) appear to correctly reflect the data weighting description in the draft report. No SAS codes can be found in appendix P for equations 3 and 4.	
Reviewer #2	Yes.	
Reviewer #3	This reviewer cannot provide details on SAS codes and Access model since he has experience in risk assessment and decision analysis and does not have advanced knowledge in SAS and Access.	
Reviewer #4	I do not have sufficient expertise in M/S Access to be able to make informed comment about the accuracy of the code underlying the RRM interface, nor whether the code accurately reproduces the logic of the model as described in Section 4 of the draft report. However, see response to Question 7b that discusses assessment of the correctness of implementation of Equation 3 and Equation 4.	
Reviewer #5	[To be completed]	

CHARGE QUESTION 8: In the Access Model, is the underlying relational database including lookup tables and algorithm appropriately designed and implemented? If not, please explain what changes should be considered.		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	The underlying relational database, look-up tables, and algorithm appear to be appropriately designed and implemented.	
Reviewer #2	No comments.	
Reviewer #3	This reviewer cannot provide details on SAS codes and Access model since he has experience in risk assessment and decision analysis and does not have advanced knowledge in SAS and Access.	
Reviewer #4	I do not have sufficient expertise in M/S Access to be able to make informed comment about the accuracy of the code underlying the RRM interface as it applies to analysis of the data in the look-up tables and whether that has been implemented as described in the draft report.	
Reviewer #5	[To be completed]	

CHARGE QUESTION 9: <i>Is the user interface of the Access Model sufficiently described for the user to understand each component of the model, e.g., foods, hazards, ranking criteria, results, and cited references?</i>		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	The user interface of the model is sufficiently described for this reviewer to understand each component of the model. However, it should be noted that this reviewer is familiar with the model approach so a newcomer may have a different opinion.	
Reviewer #2	Yes. But I get an error message when I run the ranking. But when I update the outputs I get the results. When exporting the results to Excel, all the information about the scenario is exported but not the scores and ranking outputs. So I just copy and paste from the window interface.	
Reviewer #3	This reviewer cannot provide details on SAS codes and Access model since he has experience in risk assessment and decision analysis and does not have advanced knowledge in SAS and Access.	
Reviewer #4	<p>The User Guide still refers to the model as the HRF, despite noting (in the Introduction) that the name for the model is now the Risk Ranking Model for Product Tracing (RRM-PT). This is a trivial point, but it is unnecessarily confusing, particularly given that document could have been made 'up to date' with a few minutes of 'search and replace' in Microsoft Word.</p> <p>I tried to read the User-Guide and use the RRM-PT model interface as though I <i>hadn't</i> spent many days going through the draft report in detail, i.e., to consider both from the perspective of an uninitiated potential user. From that I consider that the user-interface is not sufficiently well-described for a user to be able to use the model appropriately and to be able to correctly interpret the outputs.</p> <p>The 'User Guide (short version)' is necessary, but even that is insufficient because it assumes that the reader/user has an understanding of the structure and logic of the risk ranking model. The User Guide (Introduction) refers to the quite detailed Sections 2 and 4 of the Draft Report. The user "help" functions in the interface itself are useful, but not of themselves adequate to enable a user to correctly use the model and to interpret its outputs. It would improve the User's Guide if a few paragraphs were added to the Introduction that explained the basis of the seven criteria, the scoring scales used (including some mention of underlying data sources), how they are combined to generate a score for each food:hazard pair, and how those scores are aggregated to generate an overall risk score. The use of terms like 'scenario', 'criteria', 'repository', 'weights' etc. needs to be explained in terms of their use in the RRM-PT.</p> <p>In summary, as someone who (now) understands the basis of the RRM and its supporting data and algorithms, I found the user manual very useful to enable me to begin to use the model to generate risk ranking estimates. Without the User's Guide Short Version I don't think I would have been able to do that intuitively, even having thoroughly read and worked through the draft report, suggesting that the user interface is not 'stand-alone' for novice users.</p>	

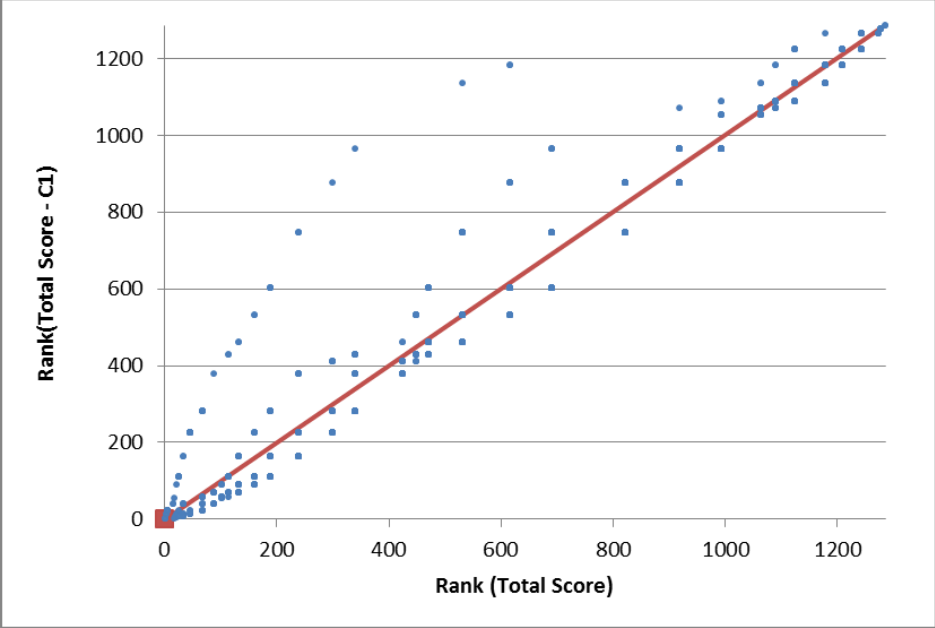
CHARGE QUESTION 9: <i>Is the user interface of the Access Model sufficiently described for the user to understand each component of the model, e.g., foods, hazards, ranking criteria, results, and cited references?</i>		
REVIEWER	COMMENT	RESPONSE
	<p><i>Other comments on the interface and user guide:</i></p> <p>The RRM-PT interface suggests that each criterion value can be traced to a data source or expert opinion source. I followed many of the ‘literature’ links and the resulting screen ‘dialog-box’ contained no information. This requires some explanation in the user guide and the interface itself. Also, sometimes a ‘literature’ hyper-link led only to description of the data as ‘expert opinion’. This aspect seems inconsistent and conveys the sense that RRM is still incomplete.</p> <p>The User Guide refers to Table 1, but Table 1 is apparently not included in the Users Guide. P3 (User Guide), third dot point about “View Baseline Data’. I think the intended word is ‘underpinning’ not ‘underlining’?</p> <p>In the dialog box about (re)assigning scores to composite score criteria, the help function doesn’t explain that a password is required, nor how to obtain it.</p> <p>P8 (User Guide) dot point 4, second line.. There is an error in the text “we be used...”, but I’m not sure what the intended text is.</p>	
Reviewer #5	[To be completed]	

CHARGE QUESTION 10: <i>How often should the model be updated, considering the data sources and data currently available and types of data that might become available in the future?</i>		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	The frequency of updates of the model should be determined by the underlying data upon which the scores for the seven criteria are developed (e.g., CDC surveillance data release, TDS monitoring data release, NHANES data release, etc.). A 2- to 4-year period for update may be reasonable.	
Reviewer #2	To answer this question, I need to know the robustness of the ranking. This can be done by moving for each food-hazard combinations the scores and see when the rank is significantly different. For the non-robust ranks it is needed to look at uncertainty scores. If the uncertainty score is high and the rank is not robust that means the score of those food-hazard combinations needs to be updated as soon as possible.	
Reviewer #3	[The reviewer did not comment.]	
Reviewer #4	The following is an extension of comments from my complementary review and report on the data behind the RR model:	

CHARGE QUESTION 10: <i>How often should the model be updated, considering the data sources and data currently available and types of data that might become available in the future?</i>		
REVIEWER	COMMENT	RESPONSE
	I am not sufficiently expert to be able to offer a reliable response to this question but would suggest that the rate of change of food processes and product formulation might mean that the model and Criterion values would need to be reviewed every 3 to 5 years. However, if it becomes evident that radically different processes or products are introduced, or products are sourced from new/different suppliers, it would be prudent to evaluate <i>before introduction of those products</i> whether those changes introduce a different level of public health risk. Nonetheless, the evaluation of the model presented here suggests that the current draft RRM is not yet ready for practical use and risk management decision-making.	
Reviewer #5	[To be completed]	

CHARGE QUESTION 11: <i>Is the draft report clear in its description of the risk ranking approach, criteria and scoring definitions, and model limitations? If not, please identify which aspects are unclear or could be more transparent.</i>		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	This is a complex process with many layers of data aggregation. The information provided in the draft report and associated appendices are well laid out and easy to followed with adequate details in most areas. The Figure 2-1 demonstrating the relationship between the FSMA factors and the criteria in the high risk food model (HRFM) is helpful to orient the model in context of FSMA. The descriptions and scoring of the seven criteria as described in Section 2 of the current draft is mostly clear and reflective of what the current model intends to do. Section 4.1 and data indicator in Table 4-2 are useful, allowing for a quick understanding of underlying data/metric that are relied upon as proxy for exposure and risk for the food-hazard pairs. The scoring process flow charts for each criterion in the HRFM supplement the description of the seven criteria in Section 2 and further clarify the current model. The descriptions of the expert elicitation and information in appendix K are helpful allowing readers a better understanding of the sources and limitations associated with the information obtained from this process to fill in data gaps. The description of risk and uncertainty scores is easily understood. Overall the report did a very good job of explaining what has been done and the elements of the model.	
Reviewer #2	In general, the report provides the needed information to understand the risk ranking process.	
Reviewer #3	The report is rather comprehensive and appears to be clearly written. It has an excellent and logical structure. Assumptions are clearly stated, and Appendices are very useful in evaluating data and calculation algorithms. The main concern is about more clarity on the meaning of the scoring rules.	
Reviewer #4	Comments about the transparency of the process and the model, identification and documentation of relevant and authoritative sources, and limitations in the data and the model were made above and also in the complementary report and review of the data. Specific comments are presented in Section III, below.	

CHARGE QUESTION 11: <i>Is the draft report clear in its description of the risk ranking approach, criteria and scoring definitions, and model limitations? If not, please identify which aspects are unclear or could be more transparent.</i>		
REVIEWER	COMMENT	RESPONSE
	Further, there appear to be many minor presentation errors and examples of use of jargon and idioms that may not have unambiguous meaning to all readers.	
Reviewer #5	[To be completed]	

CHARGE QUESTION 12: <i>Do you have any additional comments? Please share them in your review.</i>		
REVIEWER	COMMENT	RESPONSE
Reviewer #1	None.	
Reviewer #2	<p>The sensitivity analysis should not be restricted only to the choices of criteria weights. A very simple sensitivity analysis can be made by ranking the food-hazard combination with and without each of the seven criteria and see if there are some significant changes in the rank order of food-hazard combinations. As an example Figure 2 shows the changes of rank order when C1 is not included.</p> 	

CHARGE QUESTION 12: Do you have any additional comments? Please share them in your review.		
REVIEWER	COMMENT	RESPONSE
	<p>Figure 2. Impact of C1</p> <p>To evaluate the overall importance of factor C1 on the risk ranking outputs, a statistic such as The Kendall's tau rank correlation coefficient can be used.</p> <p>Moreover for each food-hazard combination it will be useful to determine if a single score can modify significantly the risk rank order: this can be done by changing the score by zero only for the considering food-hazard combination and calculating the distance between the rank with all the scores and the one obtained when one score is excluded.</p>	
Reviewer #3	<p>Additional clarification for Question 3.</p> <p>The proposed approach would be something like this: For food-pathogen pair i risks r_{ij} are defined verbally as relating to multiplicative factors such as probabilities of occurrence, magnitude of loss per given occurrence etc. so that total expected loss to a person eating a serving of the food is $r_i = r_{i1} * r_{i2} * \dots * r_{in}$. This approach uses an additive risk score and calculating total risk $x_i = x_{i1} + x_{i2} + \dots + x_{in}$. Ideally, this score works such that $x_1 > x_2$ if and only if $r_1 > r_2$ - i.e., you would never classify food A as high risk and B as low risk when the expected loss of A is lower than the expected loss of B. The scoring rules where x_{ij} are given scores from 1 to 9 will only work if x_i is proportional to $\ln(r_i)$. It is not at all clear this is the case. Visual inspection suggests that x_i have an approximately linear relationship with r_i. Even if x_i are appropriate, it does not make sense here to use weights, since if we take exponentials to convert x to r, coefficients on the x terms transform to powers on the r terms (each of which ought to be linear in expected loss).</p> <p>When considering risk for a food across multiple food pathogen pairs, we cannot simply add their scores. Instead, we would have to calculate $x_{12} = \ln[\exp(x_1) + \exp(x_2)]$. Similarly, when considering risks for multiple similar foods (e.g., tuna and salmon), we cannot simply sum them. First, we want to weight by proportion - $r_{fish} = \text{proportion_tuna} * r_{tuna} + \text{proportion_salmon} * r_{salmon}$, or more generally, if food group i consists of subgroups k, $r_i = \sum \text{over } k \text{ of } w_{ik} * r_{ik}$. To convert these back to total scores consistent with the kind of x-values I have described for multiplicative risk factors, we would need to have $x_i = \ln[\sum(w_{ik} * r_{ik})]$, and since there is no neat form for $\ln(a+b)$, the whole thing becomes kind of unwieldy if done right.</p> <p>In sum, this represents a potentially serious methodological problem that must be directly addressed. Although algorithms would not be difficult to modify, they would be less transparent for non-expert users.</p> <p>The overall approach might be solidified by:</p>	

CHARGE QUESTION 12: Do you have any additional comments? Please share them in your review.		
REVIEWER	COMMENT	RESPONSE
	(6) clarifying whether factor scores are meant to indicate magnitude or order-of-magnitude of probabilities/consequences (7) ideally they would represent order of magnitudes (8) the method would be used as-is to produce scores for the food-hazard pair (9) to calculate collective risk over hazards we would sum the exponential of the hazards and then take the logarithm of that total to calculate risk for a food group we would calculate a weighted sum of the exponential of scores for each constituent food and take the logarithm of that total.	
Reviewer #4	No. All relevant comments have been made above.	
Reviewer #5	[To be completed]	

III. Specific Observations on Draft Report for Peer Review: Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA (RRM-PT Draft Report) within the context of the model itself				
REVIEWER	Page	Paragraph/ Line	Comment	RESPONSE
Reviewer #1	17	Line 19	Table 2-1 title should reflect acute exposure	
	18	Line 8	Text in [Table 2-2] describing score 9 for chronic health hazards should include examples of endpoints with this type of score	
	63	Line 1	Examples did not show apple juice and arsenic example	
	634	Line 36	Salad kit example of no data and expert opinion was used to assign score, who are the experts? Was this from the elicitation process outline in appendix N?	
	71	Lines 12-13	Who are the allergen experts?	
	72	Lines 15-17	Who are the FDA subject matter experts? How is this done, is there a report detailing his process? An appendix documenting this process (similar to appendix K for the expert panel elicitation process) would be helpful.	
	73	Line 10	Why sum? Need to provide rationale	
	102	Line 27	Appendix L was not provided for peer review	
Reviewer #2			No comments.	
Reviewer #3			No comments.	

III. Specific Observations on Draft Report for Peer Review: Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA (RRM-PT Draft Report) within the context of the model itself				
REVIEWER	Page	Paragraph/ Line	Comment	RESPONSE
Reviewer #4	General		The phrase “in order to” can in almost all circumstances be reduced to “to” without any loss of inference or meaning. “In order to” is frequently used in the document and can be simplified to “to”	
	P13	L1	‘scoring’ should be ‘score’	
	P15	L9	‘in’ should be ‘for’	
	P15	L18	‘issue’ should be ‘issues’	
	P16	L2	‘outbreak’ should be ‘outbreaks’	
	P16	L3	‘representing’ should be ‘represent’	
	P17	L16	‘definition’ should be ‘definitions’	
	P19	L5	delete ‘or’	
	P19	L37	delete ‘foods’ after ‘ready-to-eat’	
	P19-P20	L40-L2	This sentence essentially repeats the previous sentence (i.e., P19, L36 – 39).	
	P20	L12	font size is inconsistent	
	P21	L24	primary production infers, ‘on the farm’, i.e. ‘on the farm’ is a tautology.	
	P24	L8	delete ‘be’	
	P25	L5	‘pair’ should be ‘pairs’	
	P25	L18	‘examples’ should be ‘example’	
	P26	L2	delete second ‘in’	
	P26	L13	change ‘identified’ to ‘identify’	
	P28	L6	‘polynuclear’ should be ‘polycyclic’	
	P30	last para in right-hand column	delete ‘and’ in first line of para	
	P31	L6	correct to ‘to assign a numerical score of...’	
P31	L10	correct ‘qualitatively’ to ‘qualitative’		
P33	L5	change ‘detail’ to ‘detailed’		
P37	L5	insert ‘that’ after ‘address’; change ‘pair’ to ‘pairs’		
P40	L12	change ‘hazard’ to ‘hazards’		
P44	Figure 4-3	In the second table in the right-hand column, the instructions should say “For each prevalence study, assign the data weight”, i.e., not assign the geographic weight.		
P47	L26	delete ‘to’		
P48	L1	‘there’ should be ‘those’		

III. Specific Observations on Draft Report for Peer Review: Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA (RRM-PT Draft Report) within the context of the model itself				
REVIEWER	Page	Paragraph/ Line	Comment	RESPONSE
	P48	L17	the phrase ‘did not have non-zero quantitative prevalence...’ is convoluted and I’m still not exactly sure what it means. Try to reword more clearly.	
	P48	L30	‘detection’ should be ‘detections’	
	P51	L6	insert ‘is’ after ‘contamination’	
	P52	L17	insert ‘and’ before ‘chemical hazards...’	
	P56	L27	change ‘step’ to ‘steps’	
	P64	L6	change ‘is’ to ‘as’, or delete ‘is’	
	P72	L25	FDA has expert onions? (Perhaps change to ‘opinions’?)	
	P77	L6	insert ‘the’ after ‘run’	
	P77	L29	the data don’t underline the model, they ‘underpin’ it	
	P84	L12	delete “number of”	
	P91	L19	change ‘give’ to ‘gives’	
	P96	L10	change ‘frequently’ to ‘frequent’	
	P96	L13	change ‘precipitate’ to ‘precipitous’; However, either term is a subjective expression that seems largely to describe a natural consequence of the scoring scheme, more than any real phenomenon related to food-hazard pairs. See also relevant discussion about this “observation” in response to Question 4.	
	P103	L13	insert ‘in’ before ‘Table 7-4’	
	P103	L19	insert ‘; with’ before ‘highly ranked..’	
	P103	L29 (twice), L30, L31, L39	change ‘has’ to ‘have’	
	P104	L2, L5	change ‘has’ to ‘have’	
	P107	L14	delete ‘a’ before ‘9’	
	P107	L22	insert ‘on’ before ‘option 4.’	
	P107	L26	insert ‘on’ at end of line after ‘analysis’	
	P110-111	L1-L37	the term ‘a greater degree of’ can more easily be expressed as ‘more’, i.e., there were more data gaps....	
	P111	L20	change ‘chose’ to ‘choose’	
	P122	L4 ff.	The bibliographic details are incomplete. Minor et al (2015) is now published, in Risk Analysis, 35, pp. 1135-1139.	
Reviewer #5			[To be completed]	

IV. Specific Observations on Appendices to the Draft Report for Peer Review: Risk Ranking Model for Product Tracing as Required by Section 204 of FSMA (RRM-PT Draft Report) within the context of the model itself					
REVIEWER	Appendix	Page/ Row	Paragraph/ Line/Column	Comment	RESPONSE
Reviewer #1	K	5	Second paragraph, lines 1 and 2	It is stated that criterion 7 (economic impact) has relatively few data gaps, yet looking at Table 3 on page 6, we see 340 for chemicals. 340 does not appear to be minor data gap here. If the 340 C7 scores for chemicals were based on the expert elicitation as indicated in Table 3 of Appendix K, then who were the experts from which the economic scores elicited from? On the list of experts in Table 2 of appendix K, there is no “health economics” expertise that would be necessary for such an expert elicitation.	
	N	1, Table 1	Leafy green has FRS of 249	This 249 was said to be based on sum of individual hazard pairs for leafy green in this appendix. However, in draft report on page 102, lines 28-31, it is said to have 12 pairs and the risk scores for the pairs range from 130-450. How is it possible that the sum for 12 pairs be less than the score for one pair of 450?	
	O			Why are these data not included? Is there plan to include?	
Reviewer #2				No comments.	
Reviewer #3				No comments.	
Reviewer #4				See responses to specific questions	
Reviewer #5				[To be completed]	

V. Specific Observations on Risk Ranking Model for Product Tracing: User’s Guide Short Version				
REVIEWER	Page	Paragraph/ Line	Comment	RESPONSE
Reviewer #1			No comments	
Reviewer #2			No comments	
Reviewer #3			No comments	
Reviewer #4			See responses to specific questions, in particular Q. 9.	

V. Specific Observations on Risk Ranking Model for Product Tracing: User's Guide Short Version				
REVIEWER	Page	Paragraph/ Line	Comment	RESPONSE
Reviewer #5			[To be completed]	

VI. Specific Observations - Provide specific observations, corrections, or comments on the Access Database FDA's High Risk Foods (HRF) Model				
REVIEWER	Tab	Steps taken within the tab	Comment	RESPONSE
Reviewer #1			No comments	
Reviewer #2			No comments	
Reviewer #3			No comments	
Reviewer #4			See responses to specific questions, in particular Q. 9.	
Reviewer #5			[To be completed]	